UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Analysis of the contribution of alternative splicing to glioma subtype definition

Maria Teresa Proença Mendes Maia

Mestrado em Bioinformática e Biologia Computacional
Especialização em Biologia Computacional

Dissertação orientada por:
Nuno Morais, Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa
Lisete Sousa, Faculdade de Ciências, Universidade de Lisboa

2016

RESUMO

Gliomas são tumores cerebrais que têm origem em dois tipos de células: os astrócitos e os ologodendrócitos, os quais formam uma estrutura de suporte para os neurónios. Os gliomas são responsáveis por cerca de 80 % dos casos malignos de tumor cerebral.

A classificação de gliomas fundou-se durante muito tempo em parâmetros histológicos, tais como o tipo histológico ou o estádio do tumor, uma medida do seu grau de malignidade, baseada na aparência e comportamento das células. A realização de estudos de larga escala fazendo uso das novas tecnologias ómicas, tem permitido melhorar os sistemas de classificação clássicos, através da identificação de assinaturas moleculares subjacentes aos diferentes subtipos tumorais. No caso de gliomas, uma publicação recente fez a descrição de um sistema de classificação robusto, baseado num painel de 1300 marcadores de metilação de DNA, aplicável a tumores dos estádios de 2 a 4, em que são definidos seis subtipos (LGm1 a LGm6) que formam grupos prognóstico bastante homogéneos (Ceccarelli et al., 2016).

O *splicing* é um mecanismo de processamento pós-transcricional através do qual certos segmentos de uma molécula de pre-RNA mensageiro (pre-mRNA): os intrões, são eliminados, resultando num mRNA maduro constituído por segmentos chamados exões que codifica para uma proteína (ou produto génico). É um processo químico levado a cabo por um complexo macromolecular modular, que se designa por spliceossoma. O *splicing* alternativo consiste na produção de mais do que um tipo de mRNA maduro a partir do mesmo gene através da retenção ou eliminação seletiva de um exão/intrão, dito alternativo ou regulado. Este processo contribui para a geração de diversidade funcional de produtos génicos e é regulado de forma específica em cada tecido e estádio de desenvolvimento, sendo que alterações aos seus padrões normais estão descritos como podendo promover ou apoiar o processo de tumorigénese.

A quantificação de *splicing* alternativo pode ser feita usando uma medida que se designa por index da percentagem de splicing ou PSI, e que corresponde à proporção de transcritos que incluem um exão regulado em relação ao total de transcritos de um gene.

O presente projeto de tese visa analisar a contribuição da regulação de *splicing* alternativo para a definição da classificação dos gliomas de estádios 2 a 4, tendo como objeto de estudo o conjunto de dados de uma coorte de 674 casos de glioma depositado no portal TCGA (*The Cancer Genome Atlas*).

Por forma a avaliar a existência de uma assinatura molecular própria ou associada aos subtipos de glioma estabelecidos, utilizou-se análise multivariada dos dados de quantificação de *splicing* alternativo, mas também de expressão génica. Utilizando análise de componentes principais (PCA), o *splicing* alternativo mostrou capturar diversidade biológica de forma muito semelhante à expressão génica. A componente principal associada aos dois níveis de dados transcriptómicos de maior relevância representou um gradiente de malignidade tumoral. O *splicing* alternativo demonstrou ser informativo relativamente à distinção dos subtipos LGm2, LGm3 e LGm4/5, enquanto os subtipos LGm1 e LGm6 revelaram uma grande heterogeneidade.

Análise de expressão génica e *splicing* alternativo diferencial ao longo dos subtipos LGm permitiu identificar um grupo de 5970 genes e 1762 eventos de *splicing* associados à definição desses subtipos. De forma importante, 183 genes e 105 eventos de *splicing* com regulação diferencial afetam genes cujas mutações têm implicação causal em cancro demonstrada. Por fim, 41 fatores de *splicing* apresentaram de igual modo expressão génica diferencial entre subtipos, com os genes

IGF2BP2 e IGF2BP3 apresentando os resultados mais significativos, nomeadamente uma expressão elevada em LGm1,4,5 e 6, os subtipos associado a um pior prognóstico.

Análise de enriquecimento funcional realizada com a informação de regulação diferencial da expressão génica e *splicing* alternativo entre subtipos LGm revelou funções bilógicas distintas por cada processo. Enquanto os genes com alterações de expressão entre grupos de metilação de DNA se relacionaram com funções como resposta imune, proliferação, sobrevivência e adesão celulares, genes tendo o seu *splicing* alternativo alterado involveram sobretudo o processamento de RNA, síntese proteica e também apoptose.

O valor do *splicing* alternativo e da expressão génica para o prognóstico em gliomas foi avaliado usado modelos de regressão de Cox para a sobrevida do paciente em função de diferentes fatores de risco. Um teste inicial sobre a capacidade da componente principal associada à malignidade para explicar a evolução do tempo de sobrevida do doente confirmou a superioridade desta dimensão dos dados de transcriptómica relativamente à variável estádio do tumor no que diz respeito a essa previsão. Subsequente análise de eventos de *splicing* e genes individuais como preditores de prognóstico resultou na descoberta de tantos quantos 11794 genes e 6657 eventos de *splicing*. Porém, apenas um gene e dois eventos de *splicing* alternativo foram capazes de melhorar a estimativa da evolução do doente relativamente a um modelo já contendo subtipos LGm, estádio do tumor e idade do doente, três covariáveis relevantes descritas na literatura. Os dois eventos em questão apresentaram distribuições de PSIs com variância muito baixa, pelo que constituiriam marcadores de prognóstico de pouca qualidade, além de não parecerem ter um interesse intrínseco já que não representam a possibilidade de geração de uma transição de uma isoforma dominante para outra. Finalmente, marcadores especificamente associados aos grupos LGm foram identificados a partir da interseção do conjunto que apresentou valor prognóstico independente do estádio do tumor e da idade do doente com o conjunto com regulação diferencial entre os seis subtipos. Desta análise resultou um total de 337 eventos de *splicing* alternativo, 50 dos quais acrescentando informação prognóstica relativamente aos dados de expressão génica, e também 20 genes de fatores de *splicing.* De entre estes últimos, seis codificavam para proteínas que se ligam ao RNA (RBPs) com motivos de ligação conhecidos, pelo que o seu potencial papel regulatório foi investigado.

Uma metodologia para a descoberta de mecanismos de regulação de *splicing* alternativo em *trans* foi implementada. Concretamente, um algoritmo para a geração de mapas de regulação de *splicing* específicos de cada RBP foi usado, tendo como objetivo determinar as regiões regulatórias para fomento ou silenciamento do *splicing* de exões alternativos. A identificação da posição relativa dos alvos de regulação de cada RBP baseou-se na deteção dos eventos de splicing cujas percentagens de inclusão do exão alternativo se correlacionavam com a abundância da RBP e que efetivamente continham motivos de ligação para essa RBP nas regiões vizinhas do exão regulado. Este método foi validado para o fator de *splicing* bem estudado PTBP1, utilizando tanto um conjunto de dados provindo de tecidos saudáveis como com o da cohorte de glioma estudada. No entanto, a aplicação do mesmo procedimento a quatro dos seis fatores de *splicing* potencialmente associados com as alterações de *splicing* entre subtipos de glioma resultou em mapas de *splicing* de RNA inconsistentes entre os dois conjuntos de dados. Medidas para a melhoria desta metodologia foram identificadas e poderão ser aplicadas futuramente por forma a poder concluir sobre a relevância destas proteínas em glioma.

Este estudo permitiu identificar um número de eventos de *splicing* alternativo e genes expressos, nomeadamente, genes de fatores de *splicing*, que apresentam comportamento diferencial em termos de malignidade e subtipo epigenático de glioma e que poderão ter valor diagnóstico e

terapêutico interessante. Adicionalmente, um novo método computacional para descoberta de mecanismos de regulação de *splicing* alternativo foi implementado, tendo permitido propor um mecanismo de ação para o fator de *splicing* relevante em glioma KHDRBS2.

Palavras-chave: *Splicing* alternativo, Glioma, Cancro, Transcriptómica

ABSTRACT

Gliomas are brain primary tumours that originate from two kinds of glial cells: astrocytes and oligodendrocytes, which make up a supportive structure for neurons.

Classification of gliomas has for long relied on histological parameters, like cell type composition and grade, a measurement of the degree of malignancy of a tumour based on cell appearance and behaviour. Large scale studies employing the new omics technologies have allowed to improve classic classification systems, through the identification of molecular signatures behind each tumour subtype. In the case of gliomas, a recent publication described a robust classification system based on DNA-methylation profiling, applicable to tumours from grades 2 to 4 and defining six subtypes (LGm1 to LGm6) forming quite homogeneous prognostic groups (Ceccarelli et al., 2016).

Alternative splicing is a post-transcriptional mechanism of regulation of gene expression that contributes to generate functional diversity of gene products through the selective elimination of certain segments of pre-messenger RNA (pre-mRNA) molecules. Alternative splicing is regulated in a tissue and developmental specific way and alterations to its normal patterns have been extensively reported to promote or help sustaining tumorigenesis.

The work presented here aimed at determining the contribution of alternative splicing to glioma subtype definition, having as a focus a cohort of 674 cases of glioma grades 2 to 4, coming from the Cancer Genome Atlas (TCGA) data portal.

Differential gene expression and differential splicing analyses across LGm subtypes allowed to identify a group of 5970 genes and 1762 events of alternative splicing whose regulation is associated with subtype definition. Importantly, among these differentially regulated markers, there were 41 splicing factor genes and 46 splicing factor gene-associated events of splicing.

In order to enquire about the existence of particular molecular signatures in glioma, multivariate exploratory data analysis was performed on alternative splicing and also gene expression data. Alternative splicing showed to capture sample diversity in a way that was very similar to gene expression. The most revealing principal component associated with both transcriptomic data levels presented a gradient of tumour malignancy. As for the ability of alternative splicing to distinguish subtypes, it could partially separate LGm2, LGm3 and LGm4/5 groups, while LGm1 and LGm6 revealed a high heterogeneity.

The value of alternative splicing and gene expression in glioma prognosis was evaluated using Cox regression models for patient's overall survival outcome as a function of different predictors. An initial test on the ability of the malignancy associated principle component to explain patient outcome strikingly confirmed the superiority of this dimension of transcriptomic data to make this prediction in relation to tumour grade. In turn, analysis of individual alternative splicing events and expressed genes as prognosis predictors resulted in as many as 11794 genes and 6657 events of splicing. However, only one gene and two alternative splicing events were able to improve patient survival outcome estimation relative to a model that already accounted for LGm subtype, tumour grade and patient age, three relevant covariates described in the literature. Finally, in terms of prognostic markers specifically associated with LGm groups, a total of 337 splicing events were found, 50 of which adding information in relation to gene expression, and also 20 splicing factor genes. From these latter, six encoded RNA-binding proteins (RBPs) with known RNA-binding motifs and their potential regulatory role was investigated.

A methodology for the discovery of mechanisms of alternative splicing regulation in *trans* was implemented. Specifically, an algorithm to generate maps of splicing regulation specific of each RBP was used, aimed at determining regulatory regions for their enhancing or silencing role in splicing of alternative exons. This method was validated for the known splicing factor PTBP1, both using an RNA-seq dataset from healthy tissues and the studied glioma one. However, application of the same procedure to four of the six splicing factors potentially associated with alternative splicing changes across glioma subtypes resulted in RNA splicing maps that were inconsistent between the two datasets. Improvements to the methodology used were identified and may be applied in the future so that stronger conclusions about the relevance of these proteins in glioma can be taken.

This study allowed to outline a number of alternative splicing events and expressed genes, namely splicing factor genes, that behave differently according to glioma malignancy and epigenetic groups and that may be of interesting diagnostic and therapeutic value. Also, a novel computational method for discovery of mechanisms of regulation of alternative splicing was implemented and allowed to proposal a mechanism of action for the glioma-relevant splicing factor KHDRBS2.

Keywords: Alternative splicing, Glioma, Cancer, Transcriptomics

ACKNOWLEDGMENTS

I would like to start by thanking Nuno for welcoming me to his lab to accomplish this final, very important part of my training as a bioinformatician. You have been a great supervisor, having made a very important contribution to this project, taught me a whole lot in statistics and quantitative methodologies, and the way to use these to tackle biological questions. Thank you as well for your contribution to this manuscript.

I would like to thank professor Lisete Sousa for accepting to co-supervise me and for always being available to offer guidance throughout this work. Your help has been precious. Thank you very much for taking so much of your time to help in the preparation of this manuscript.

I would like to thank in advance the members of the jury of my thesis defence for reading carefully this work and participating of my getting this degree.

I would like to thank my colleagues Marie, Lina, Mariana, Bárbara, Nuno, Carolina, Juan and Bernardo, who make up this fresh, creative, talented and very friendly Computational Biology team. I have enjoyed very much meeting you all and, believe me, to share the workspace, even if I ended up seating on the other side of the room. I always enjoyed our open discussions, exchange of thoughts about whatever subject and mutual support. I learned difficult concepts and interesting tricks with each of you.

I would like to thank Cláudia Faria for her invaluable insights, specially while kick-starting this project in brain tumourigenesis.

I would like to thank Ana Rita Grosso for her availability to give me guidance in some moments and for always being ready to share her knowledge with us.

I would like to thank all the people at IMM. It has been a great environment to work.

I would like to thank my friends, of whom I'm very proud and that have made such great companions along the years.

Thank you Florent, for sharing your life with me. Thank you for all your support. Your help in making me get through doubts every now and then while writing this thesis was also precious.

A big thank you to my parents and brother, who have always been so encouraging and supportive, including of my career decisions and to whom I owe very much. Thank you for being there.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A3 – Alternative 3’ splice site

A5 – Alternative 5’ splice site

AF – Alternative first exon

AL – Alternative last exon

AS – Alternative Splicing

bp – base pair

CRAN - The Comprehensive R Archive Network

DAS – Differential Alternative Splicing

DGE – Differential Gene Expression

GE – Gene expression

GTEx – Genotype-Tissue Expression program

KEGG – Kyoto Encyclopaedia of Genes and Genomes

mRNA – messenger RNA

MX – Mutually exclusive exons

PC – principal component

PCR – Polymerase Chain Reaction

RI – Retained Intron

RPKM – Reads Per Kilobase per Million mapped reads

SE – Skipping Exon

TCGA - The Cancer Genome Atlas projec

1. INTRODUCTION

In this thesis, the alternative splicing patterns of glioma, the central-nervous system (CNS) glial-cell derived tumour type, will be studied. The introductory text that follows will cover (1) overall background related to this type of tumour, and its classification system, (2) the theory behind alternative splicing mRNA processing, as well as the methodological approaches taken to be able to study splicing transcriptional output and, finally, (3) an overview over what is known in terms of the role of alternative splicing in gliomagenesis.

1.1 GLIOMA

1.1.1 The most pervasive CNS primary tumour

Glioma, without being the most frequent primary tumour affecting the brain, does account for about 80 % of the malignant cases. Indeed, data from the 2008-2012 report from the Central Brain Tumour Registry of the United States estimate that, from the 32.8 % of malignant cases, 15.1 % are glioblastomas, the most aggressive glioma type, and 11.3 % correspond to other malignant gliomas (Figure 1.1) (Ostrom et al., 2014). The remaining 1.1 % of gliomas are benign, i.e. have a slow pace, localized growth, which makes them not life threatening once diagnosed.

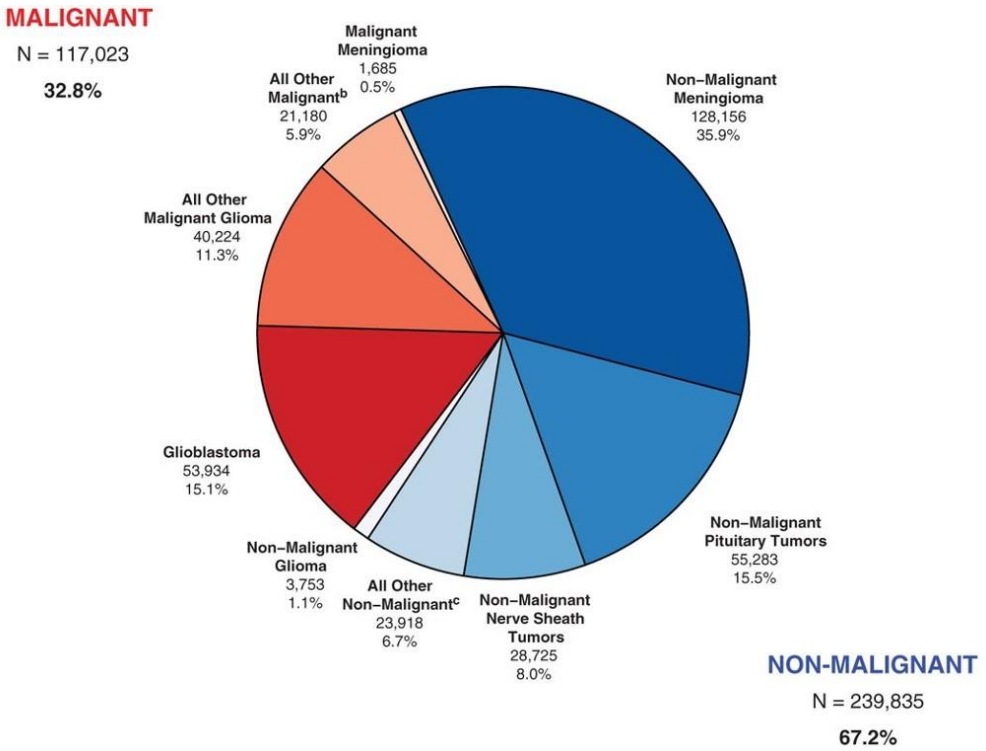


Figure 1.1 – *Distribution of Primary Brain and CNS Tumours by behaviour (N = 356,858), CBTRUS 2008-2012 Statistical Report* (Ostrom et al., 2014) .

1.1.2 Cells of origin

Cells from the glia make up a supportive structure for neurons in the brain. They can be of four essential types: astrocytes, oligodendrocytes, microglia, and ependymal cells (Snell, 2010). Astrocytes are at least as numerous as neurons and responsible for controlling ion levels and the pH at the neurological synapse, for providing nutrients to neurons (e.g. glucose and metabolic intermediates), to clear out neurotransmitters or other neuron-secreted compounds from the extracellular space, and for helping define the blood-brain barrier by means of interaction with blood-vessel endothelial cells (Nedergaard, Ransom, & Goldman, 2016). Oligodendrocytes are cells that electrically insulate neuronal processes by wrapping them around myelin sheets. These two latter cell types are also the ones that lead to glioma formation. In fact, gliomas appear as cell masses of astrocyte-like, oligodendrocyte-like or a mixture of astrocyte- and oligodendrocyte-like cells. Interestingly, these two-cell types originate from the same population of cells that also give rise to neurons, which may be called the neuroglial progenitors, and are responsible for neural/glial tissue regeneration (Modrek, Bayin, & Placantonakis, 2014).

As for the other glial cell types, which do not share a common developmental origin with the neuroglial cells, microglia are monocyte-like cells that can act like macrophages and have a neuro-protective role and ependymal cells are multi-ciliated cells that line up the brain ventricles and propel the cerebrospinal fluid.

1.1.3 Classification of Glioma

The definition of glioma subtypes is still an ongoing process, which has gained much improvement in the last years, with the contribution of studies that integrate histological and high-throughput molecular data to then relate it with patient survival data. Glioblastoma has been subjected to more thorough research before lower-grade gliomas (LGG) did and the description that follows reflects this chronological order in the evolution of glioma classifiers.

1.1.3.1 The Cancer Genome Atlas as a privileged source for cancer research

Cancer, being a complex disease, will arise as a result of genetic background and environmental exposures that are particular to the individual carrying it. As such, research on its aetiology will be complicated by the presence of “passenger” mutations or other cellular alterations carried by the individual, but that do not contribute to the development of the disease. This creates the need to carry out cancer studies in as large cohorts as possible and preferably to get access to clinical parameters that will allow not only to relate subtypes of the disease with particular cancer patient strata but also to establish a direct link between a molecular signature and the progression of the disease within an individual clinical case.

The Cancer Genome Atlas (TCGA) is a project created in 2005, as a collaboration between the US National Cancer Institute (NCI) and the US National Human Genome Research Institute (NHGRI), with the aim to create a very large source of molecular, histological and clinical data relative to more than 11000 cancer cases and 35 tumour types, all put together and made available for public use (“TCGA Home - TCGA - National Cancer Institute - Confluence Wiki,” 2016). The project had a very important impact on cancer research done worldwide, due to a very effective spread of the information. To start with, the existence of a network of researchers more directly implicated in the development of the project guaranteed the publication of the main findings gathered around multi-platform data analysis from each cancer cohort. Then, a free-access data portal was made available to the scientific

community (“The Cancer Genome Atlas - Data Portal,” 2016), with releases occurring even before publication. Access to this large amount of data, collected and analysed according to high quality standards, can be done at different levels or tiers, with tier 3 corresponding to access to clinical and processed data files, while tier 1, controlled access, includes access to all raw sequencing data (e.g. exome-sequencing, RNA-sequencing, bisulfite sequencing) and also additional information on patients’ genetic variants. Finally, cancer genomics online portals like cBioPortal (“cBioPortal for Cancer Genomics,” 2016) or COSMIC (“COSMIC: Catalogue of Somatic Mutations in Cancer - Home Page,” 2016) have been created or largely expanded through incorporation of TCGA data into their databases, which constitutes a very useful tool to be used by basic and applied researchers or by geneticists.

1.1.3.2 Glioma classification systems

Glioblastoma multiforme (GBM), a WHO grade IV tumour (Louis et al., 2007), is the most aggressive glioma type known, characterized by a high capacity to invade the surrounding tissue, high proliferation rates, abundant vascularization and a large amount of necrotic cells. It has also been characterized by the presence of numerous copy-number variants (CNV), mostly amplification of *EGFR, PDGFRA, CDK4, MDM2* and *MDM4*, amplification and or mutation of class II phosphatidylinositol 3-kinase (PI3K) genes, deletion or mutation of *PTEN, NF1, RB1, CDKN2A* and *CDKN2B* genes, and deletion of chromosome arm 10q (Brennan et al., 2013; Network, 2008). The alterations described affect three main signalling pathways relevant for cancer progression: the growth factor receptor tyrosine kinase (RTK) pathway, the p53 apoptotic pathway and the retinoblastoma (Rb) cell cycle progression pathway. In 2010, another study created a GBM subtype classifier, which consisted of four groups of tumours: Proneural, Neural, Classical and Mesenchymal, carrying mutually exclusive combinations of the previously mentioned RTK-related genomic abnormalities and, importantly, alterations in gene expression coherent with those mutations (Verhaak et al., 2010). More specifically, the Classical subtype was associated with *EGFR* overexpression, Mesenchymal subtype with *NF1* downregulation and the Proneural subtype with both *PDGFRA* overexpression and *IDH1* downregulation. The four groups described were found to correspond to similar patient prognosis. However, an interesting observation in terms of the usefulness of this GBM classifier for clinical management of patients was made: patients carrying a Classical subtype GBM were much more responsive to aggressive chemo- and radiotherapy, having improved survival times when treated, than patients carrying the Proneural subtype, which were almost unresponsive to treatment.

A new picture about the ability to define strata of patients with clearly different prognosis emerged from studies where DNA methylation profiling was carried out. Noushmehr and collaborators found that GBM and lower grade glioma (LGG) patients carrying a mutation in the *IDH1* gene had a high level of DNA methylation of their gene promoters in relation to patients carrying unaffected, wild-type copies of *IDH1* (Noushmehr et al., 2010). They termed this phenotype of high or low CpG island methylation as glioma-CpG island methylator phenotype, or G-CIMP, and found it to be more prevalent among LGG cases, in association with better prognosis.

Yet another two subtypes of GBM tumour involving epigenetic alterations have been identified (Sturm et al., 2012), which are found only in paediatric cases, each one dealing with a mutation affecting amino acids K27 or G34 of the histone H3.3.

As for LGG (grades II and III), genomic sequencing and CNV analysis has also allowed to get a good overall picture of their associated genetic lesions. One of the works that better made this description

is the work of Suzuki et al. (Suzuki et al., 2015), where three main types of grade II and III gliomas are defined and which brought as main findings, on the one hand, the presence of at least one CNV in roughly all tumour samples, the most frequent of which being a co-deletion of chromosome arms 1p and 19q, and, on the other hand, frequent mutations in the following genes: *IDH1*, *TERT* promoter, *TP53, ATRX, CIC* and *FUBP1*, with particular note going to IDH1 gene, which was estimated to be mutated in 75 % of grade II and III gliomas. The three types of LGG consisted then of: type I, carrying mutated *IDH*, a 1p/19q codeletion and *TERT* promoter mutation, with or without *CIC* and *FUBP1* mutations; type II LGG were also *IDH*-mutant, *TP53*-mutant and frequently carried *ATRX* mutations, resulting in a lower overall survival of patients in relation to type I tumours; type III LGG patients, which were the group found to have worst prognosis, closer to the one of GBM patients (29.1 % rate of 5-year survival), carried a normal copy of *IDH1* and also mutations similar to the ones found in GBM, affecting e.g. *EGFR, PDGFRA, PTEN, RB1* genes.

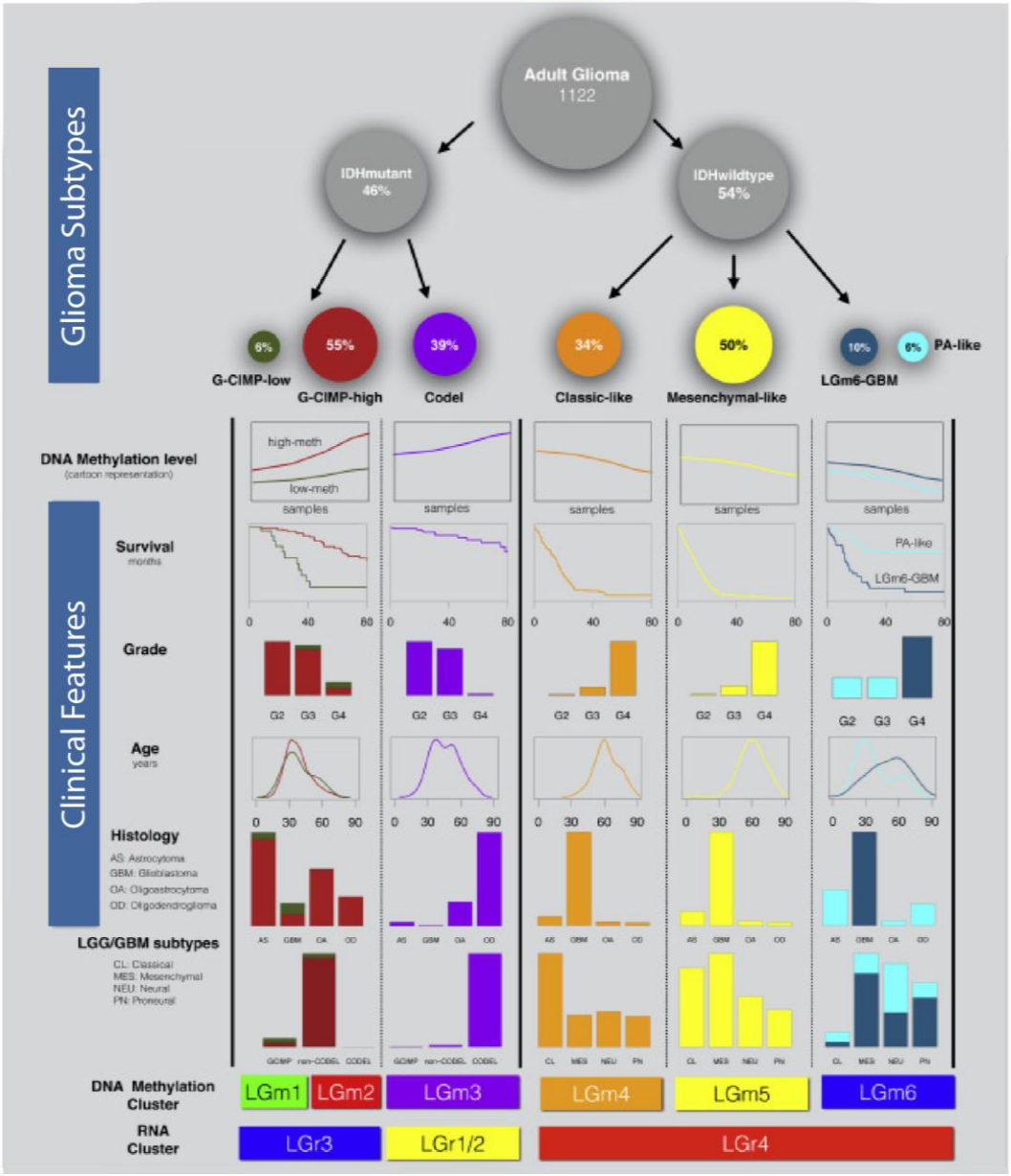


Figure 1.2 – Defining features of pan-Glioma classification proposed in (Ceccarelli et al., 2016) *and its relation with other established glioma classifications and clinical parameters. Image adapted from* (Ceccarelli et al., 2016).

Because there was a superposition of genomic lesions among GBM and LGG, it became clear the need to create a pan-glioma classifier. This was accomplished already this year, by Ceccarelli and collaborators (Ceccarelli et al., 2016), who, based on DNA-methylation profiling, were able to, with one single molecular analysis platform (bisulfide sequencing), build a pan-glioma classifier that is able to identify six main groups: LGm1 to LGm6 termed LGm DNA methylation clusters (Figure 1.2). Moreover, the authors show that this epigenetic molecular signature has prognostic value. Indeed, the authors show, using the same cohort of patients, that the power to predict patient outcome using this new DNA methylation classification, together with tumour grade and age of the patient, is superior to the one of any other classifier previously described, alone or when combined with the same two clinical variables.

A new classification of tumours affecting the CNS has been recently published by the World Health Organization (WHO), which makes use for the first time of molecular information associated to specimens, together with histological information (*see* Table S1 for a summary of the novel glioma classification) (Louis et al., 2016).

1.2 Alternative Splicing

Splicing is one of the mechanisms of messenger RNA (mRNA) processing whereby pieces of the transcribed molecule are selectively eliminated. As a result, splicing plays fundamental roles in dictating the stability of the mRNA species produced and, most of all, in deciding which protein will be generated from the mature mRNA. In fact, splicing is known to lead, within the same organism and even in the same cell, to the production of distinct transcripts and corresponding protein products, in a process that is regulated in order to meet the cell's needs in terms of protein composition. This concurrent generation of diverse transcript species from one gene through splicing is called alternative splicing (AS), which will be described in the following sections.

1.2.1 Alternative Splicing and Its Regulation

Splicing is a chemical reaction by which segments of the pre-mRNA that are not to be incorporated in the mature RNA, called introns, are extracted, while the remaining mRNA segments, called exons, are joined to remake a continuous RNA strand (Figure 1.3A). Each of these reactions can be called an event of splicing. It occurs through two trans-esterification reactions, the first one involving the 3' hydroxyl group of an adenosine residue in the intron, called the branch point, and the phosphate of the guanosine residue located at the starting position of the intron: the 5' splice site. This reaction originates the formation of a loop-like structure (lariat). A second similar reaction follows that consists on the interaction of the 3' hydroxyl group of the exon that was displaced with the phosphate group of the 3'splice site, a reaction that results in the junction of the two exons and the release of the lariat segment, which will be targeted for degradation.

Splicing is carried out by a modular protein complex, called the spliceosome, whose composition changes dynamically and, most of all, whose catalytic protein subunits responsible for carrying out the splicing reaction only assemble after the exon boundaries, the 5' splice site and 3' splice site, have been located. The recognition of the 5' and 3' splice sites is done by the U1 and U2 small nuclear ribonucleoproteins (snRNPs), respectively, which bind those sites by RNA base-pairing. Because the sequences of the splice sites are not always the same (Figure 1.3B), these are bound as

efficiently as their sequence affinity for the U1/U2 snRNPs, a condition that turns them into strong or weak splice sites.

Alternative splicing occurs in all eukaryotic cells, and is known to be abundant in organisms like plants and in human, whose 90 % of genes enable the creation of more than one transcript and protein isoform each (Pan, Shai, Lee, Frey, & Blencowe, 2008; Yang et al., 2016).

This process occurs in a tissue and developmental stage specifically regulated way and relies on the frequency with which the U1/U2 snRNPs "find" the alternative exon, that is, an exon that, unlike constitutive exons, is not always included in the mature transcripts from that particular gene.

Exon recognition is influenced by different factors, including RNA polymerase II transcriptional elongation rate, which is now established to be anti-correlated with splicing efficiency (de la Mata et al., 2003; Dujardin et al., 2014; Moehle, Braberg, Krogan, & Guthrie, 2014).

In addition, exon recognition, and thus splicing regulation, is carried out by *cis* elements, i.e. regulatory sequences located in the vicinity of the alternative exon, and by a group of *trans* acting regulators, which are proteins that alone or with interacting partners bind to the *cis* elements, to promote, or else to inhibit, the recruitment of the spliceosome machinery proteins to the alternative exon splice sites (Figure 1.3B). The *cis* elements are of four types: exonic splicing enhancers (ESEs), exonic splicing silencer (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs). The *trans* regulators are proteins that are classified as splicing factors and that may enter in the classification of RNA-binding proteins (RBPs) if they establish direct protein-RNA interactions. The two main families of alternative splicing *trans* regulators are the serine/arginine-rich proteins (SRs) and the heterogeneous nuclear ribonucleoproteins (hnRNPs). However, there are several dozens of splicing factors known, some of which are tissue-specific, whose targets events of alternative splicing have been uncovered. RBP-regulated alternative splicing events have been discovered through the study of transcript changes in gene knock-out models or RNA silencing experiments, but also by biochemical methods like cross-linking-immunoprecipitation sequencing, or CLIP-sequencing, which allow the transcriptome-wide detection of direct protein-RNA interactions. This is the case of proteins like NOVA, important during neural differentiation (Licatalosi et al., 2008), RBFOX1-3 and PTBP1, relevant during development and in adult tissues, like the brain or muscle (Y. I. Li, Sanchez-Pulido, Haerty, & Ponting, 2015)(Weyn-Vanhentenryck et al., 2014), or QK in myogenesis (Hall et al., 2013).

From these studies new ideas emerged about alternative splicing control. An important one is that alternative splicing is context dependent. Indeed, several examples showed that what determines if the splicing factor will have an activating or suppressive role on the decision to include an exon in the mature transcript is not necessarily the *cis*-element sequence, but rather the position where it stands (Figure 1.3C). As a result, it has now become an important goal for alternative splicing researchers to establish what is called an RNA-binding map for each RNA-binding splicing factor, which can be mainly accomplished by experimental designs that include CLIP-sequencing technology, but that has also been attempted *in silico*, namely through motif-enrichment approaches (Park, Jung, Rouchka, Tseng, & Xing, 2016; Paz, Kosti, Ares, Cline, & Mandel-Gutfreund, 2014). Importantly, this kind of studies have been helped by the publication of a list of RNA-binding motifs for a total of 204 RBP coding genes from 24 eukaryotic species generated by the work of Ray and collaborators (Ray et al., 2013), who through biochemical assays and computational analysis elucidated the combinations of 7-nucleotide spanning RNA nucleotides better suited for binding of each RBP.

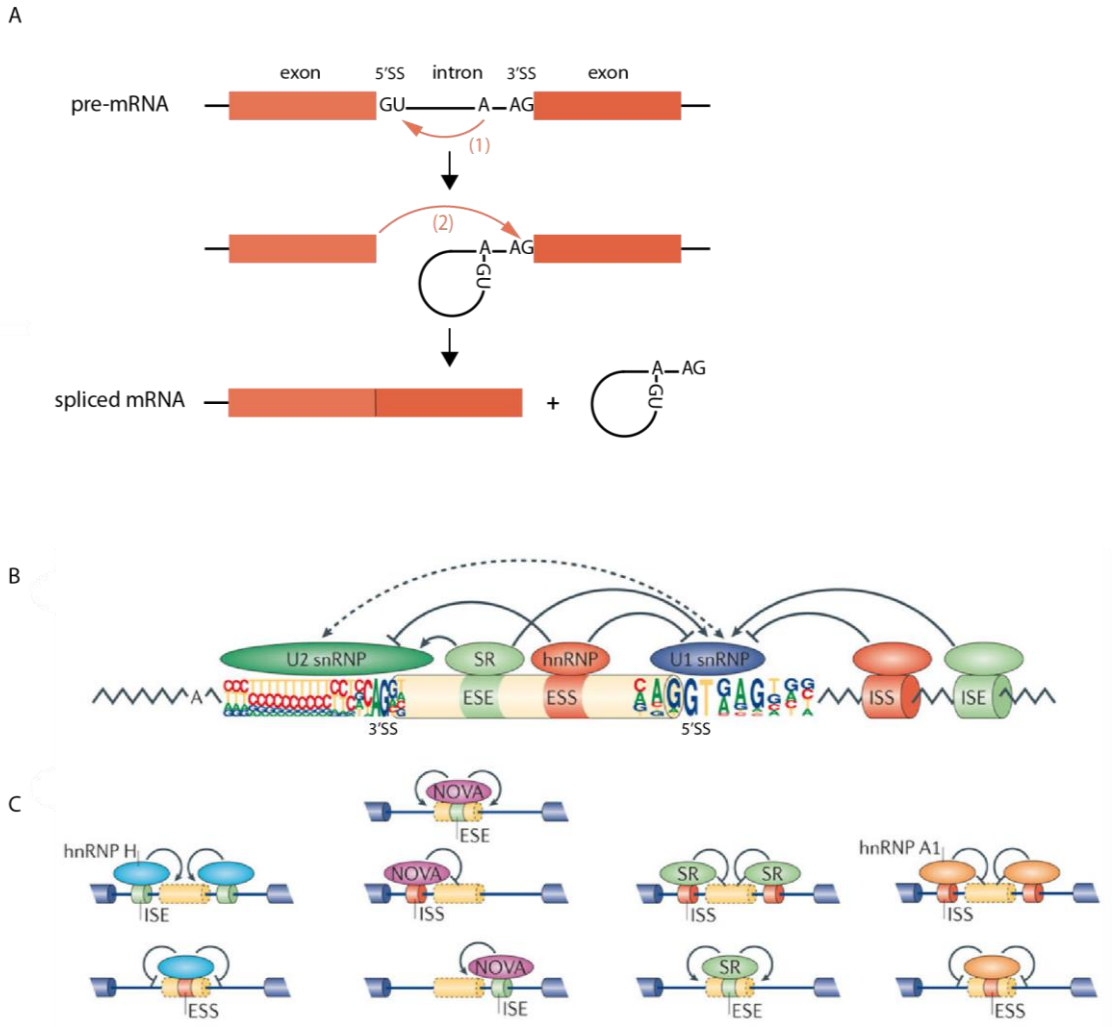


Figure 1.3 – Splicing reaction and splicing regulation. (A) Two-step pre-mRNA splicing reaction or event of splicing. (B) Splice site choice is regulated through cis-acting splicing regulatory elements (SREs) and trans-acting splicing factors. Based on their relative locations and activities, SREs can be classified as exonic or intronic splicing enhancers and silencers (ESEs, ISEs, ESSs or ISSs). These sequences recruit splicing factors to promote or inhibit recognition of nearby splice sites. Common splicing factors include SR proteins and hnRNPs, which assist U2 and U1 snRNPs during spliceosomal assembly. (C) Characterized examples of context-dependent alternative splicing regulation by SREs and four splicing factors. B and C panels of this figure are adapted from (Matera & Wang, 2014)*.)*

1.2.2 Alternative Splicing and Its Different Forms

Different types of alternative splicing events have been described, according to the relative position the included or excluded regulated exon has in relation to the competing splice sites and also the way it is annotated itself (Figure 1.4). The seven possible alternative splicing event types are: (1) skipped exon (SE), which involves an exon flanked as usual by two introns (a "cassette" exon); (2) mutually exclusive exons (MX), where a choice is made to retain either one of two "cassette" exons; (3) retained intron (RI), the possibility that the spliceosome reads-through an intron, thereby keeping it in the final transcript; (4) alternative 5' splice site (A5), in which there is competition for spliceosome recruitment between two 5' splice sites of an intron; (5) alternative 3' splice site (A3), equivalent to the 5' splice site case in which either of two competing 3' splice sites will be used; (6) alternative first exon (AF) that involves the inclusion of one out of two concurrent first exons; and (7) alternative last exon (AL), related to the inclusion of one out of two concurrent last exons. The AF type of alternative splicing is usually not directly linked to spliceosome regulation, since the choice of

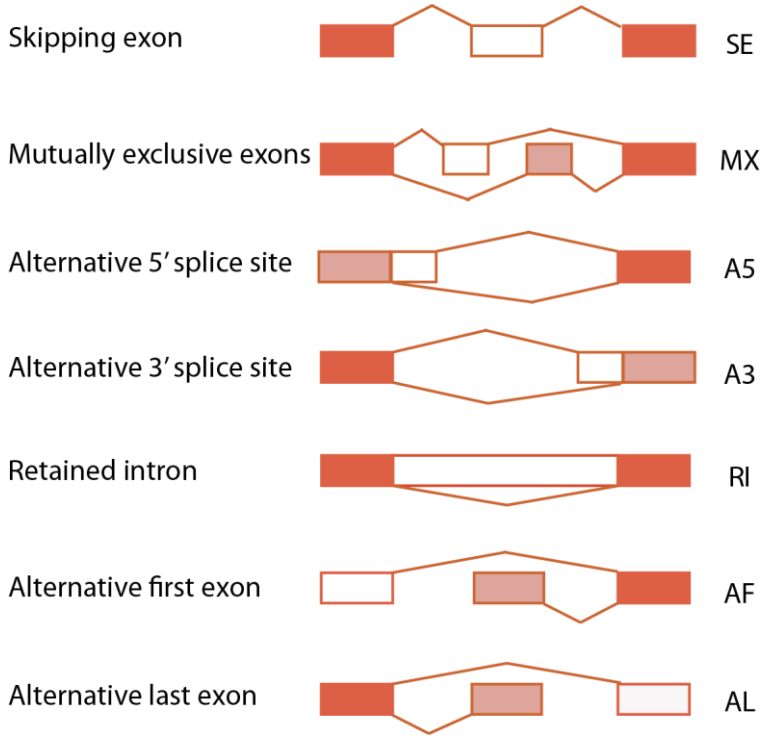


Figure 1.4 – Alternative splicing event types. White-and light rose-filled boxes represent alternative exons. InN the cases of A3 and A5 event types, the alternative white segment of the transcript forms a larger exon with the flanking exon fragment. Scheme is adapted from (Alamancos et al., 2014).

the first exon incorporation into a transcript is made through alternative transcription initiation sites.

1.2.3 Quantification of Alternative Splicing

Alternative splicing quantification can be done using as measure a relative ratio of the levels of transcripts that include a regulated exon over the levels of transcripts that do and do not include it. In this form, alternative splicing quantification assesses the rate of selection by the spliceosome of a pair of splice sites over an alternative pair.

Alternative splicing can be measured from relative abundances of transcript isoforms obtained through single gene transcripts PCR, or through high-throughput, multi-gene encompassing methodologies, like expression microarrays and RNA-sequencing. RNA-sequencing technology offers great advantages over nucleotide probe hybridization techniques because, by giving access to the actual mRNA sequence, it allows to distinguish very similar transcripts, including new ones, apart from easily providing accurate measurements of mRNA species abundance across a larger range of expression.

1.2.3.1 A note on next-generation sequencing transcriptomics

The designation next-generation sequencing refers to a group of technologies of nucleotide sequencing, namely of DNA, which, from millions of different DNA species (molecules) that are attached to a solid phase structure and physically compartmentalized, allow to follow this same number of sequencing reactions individually.

Next-generation RNA-sequencing (RNA-seq) is usually run on DNA molecules that have their complementary sequences, called complementary DNA or cDNA, and which are produced by an enzyme called reverse transcriptase that from the RNA molecule template synthesizes the corresponding DNA molecule. The initial step of an RNA-seq experiment thus consists of, from a population of total RNA or mRNA from a biological sample, produce an equivalent copy of cDNA molecules.

The most commonly used next generation sequencing technique, commercialized by the company Illumina™, is based on the sequencing by synthesis principle, and is going to be briefly described. It starts with a step of fragmentation of RNA molecules into pieces of similar sizes, under 1000 bp long, identifying all fragments from the sample with a nucleotide sequence that includes a sample barcode, oligonucleotide primer sequences and an adapter to promote the attachment of the DNA fragments to the solid state sequencing unit. This pool of DNA fragments coming from one sample is then amplified by PCR, after which it is called a sequencing library. This library is then hybridized to a flat surface called a flow cell that contains millions of binding sites for the attachment of unique DNA fragments. DNA sequencing is finally carried out in aqueous phase, using DNA polymerase enzymes and each of the four DNA nucleotides tagged with a particular fluorophore. These nucleotides have a chemical group that prevents more than one nucleotide to be added to the nascent synthesized DNA molecule at a time. Therefore, the addition of each nucleotide is followed in a controlled way by capturing a fluorescence signal, then the chemical group is released and the DNA synthesis resumed with addition of new fluorophore-tagged nucleotides.

The sequencing results come in a file that contains nucleotide sequences of each of the molecules synthesized in the flow cell, each of which will make a sequencing read, and can then be used for downstream analysis. Namely, using specialized software to carry out this analysis, reads coming from an original mRNA sample can be aligned to a reference genome, and these data can be used for quantification of exons, introns, transcripts and genes in the original sample. Quantified features are presented in individual files and are called raw counts (i.e. the raw number of reads mapping to a feature).

1.2.3.2 The Percent Splicing index

The above mentioned alternative splicing metric that expresses the relative frequencies at which the spliceosome efficiently splices an alternative exon in order to incorporate it in the mature mRNA molecule takes the name of percent splicing index, percent-spliced in, PSI or ψ (Venables et al., 2008; E. T. Wang et al., 2008). The general formula of PSI calculation for an event of alternative splicing affecting one gene, and given a group I of transcript isoforms that include the stipulated alternative exon and a group E of transcripts in which this exon is not spliced, is as follows:

$$PSI = \frac{\sum_{i \epsilon I} Number\ Inclusive\ Transcript_i}{\sum_{n \epsilon I \cup E} Number\ Transcript_n},$$

and takes values from 0 to 1.

While using RNA-seq data, the way transcript numbers are estimated varies according to the software algorithms and options used. The transcripts considered may come from a previous reference annotation of the genome or else assembly of new transcript isoforms may be allowed during the analysis. Not only that, PSI values can in practice also be calculated from numbers of sequencing reads that span the exon-exon junctions involved in the alternative splicing event, or the reads that span both the exon-exon junctions and the alternative exon individually. This way of computing PSIs is called event-centric, in contrast to the isoform-centric that departs from counts

relative to the whole transcript isoform. The work presented here makes use of the isoform-centric approach.

1.3 Alternative Splicing in Glioma

Alterations of alternative splicing patterns have been extensively reported to promote or help sustaining tumourigenesis. In 2007, a first publication showed how a change in the expression of a splicing factor could cause malignant transformation (Karni et al., 2007). In this work it was shown that fibroblasts overexpressing SRSF1 protein induced tumour formation through transplantation in a mouse model, with that overexpression producing switches in the relative abundances of oncogenic and tumour suppressive transcript isoforms that explained the malignant behaviour. More recently, already with the use of RNA-seq data, namely coming from the large cohorts of the TCGA project, pan-cancer studies have documented the existence of alternative splicing patterns that are cancer- and also cancer-type specific (Danan-Gotthold et al., 2015; Sebestyén, Zawisza, & Eyras, 2015; Tsai, Dominguez, Gomez, & Wang, 2015).

There are already descriptions of recurrent alternative splicing alterations in glioma, mainly in glioblastoma. In terms of known splicing factors implicated in this disease, there are PTBP1, PTBP2, A2BP1 (RBFOX1) and MBNL1. Although only PTBP1, and not PTBP2, gene expression alterations have been detected in tumour samples or glioma cell lines, it was shown by Cheung and collaborators that the down-regulation of these proteins in glioma cell lines had an onco-suppressive effect, reducing cell division rhythms and cell migration with a contrasting increase in cell adhesion (Cheung et al., 2009). Microarray expression analysis revealed that *PTBP1* expression reduction promoted the inclusion of exon 3 of the *RTN4* gene, thus leading to the expression a protein that reduced cell proliferation. In another study a new important target of PTBP1 splicing regulation was discovered: the tumour suppressor annexin 7 gene (*ANX7*) (Ferrarese et al., 2014). Once again, an alternative exon silencing role for PTBP1 was found in glioblastoma cells, where the inclusion of *ANX7* exon 6 transcripts was suppressed resulting in decreased targeting of the EGFR growth factor receptor for degradation. Another well-studied example of the impact of splicing factors in glioma is that of the A2BP1 protein. Usually expressed in differentiated cells from the neuronal lineage, this protein was found to be downregulated in glioblastoma, resulting in a compromised terminal differentiation and acquisition of tumorigenic properties of neural stem cells (Hu et al., 2013). In this work, TPM1, a cytoskeletal remodelling protein, was found to be a crucial target of A2BP1, whose lack of splicing contributed to the malignant transformation.

Then, many other examples of splicing events that specifically affect glioma have been studied, some of which will be described. Growth factor receptor *FGFR1* gene codes for two protein isoforms, α and β, this latter one missing exon 3 that encodes an extra NH2 extracellular loop that leads to a higher affinity towards the ligand and thereby to increased GBM cell growth (Yamada, Yamaguchi, Brown, Berger, & Morrison, 1999). *FGFR1* was found to be upregulated in glioblastoma, with concomitant switch of the prevalent FGFR1 isoform to the β form (Yamaguchi, Saya, Bruner, & Morrison, 1994). Another growth factor receptor, *EGFR*, which is the most mutated gene in glioblastoma and usually appears overexpressed in this tumour, has been shown to have splice site mutations for exons 2 and 22, although these mutations are not among the most frequent (Brennan et al., 2013).

A final interesting example of a gene whose alternative splicing ratios greatly impact on glioma patient prognostic outcome prediction is the *Reversion-inducing Cystein-rich protein with Kazal*

motifs (RECK) gene. Until recently, only one protein isoform for this gene was known, with tumour invasion, angiogenesis and metastasis suppressive properties, through the downregulation of the extracellular matrix degrading metalloproteinases MMP-9, MMP2 and MMP14. A recent study (Trombetta-Lima et al., 2015) has shown the existence of two novel isoforms for RECK: RECK-B and RECK-I. Furthermore, this study shows that patients with their high-grade gliomas having higher ratios of the RECK transcript that encodes the canonical isoform are associated with a better overall survival. Experiments performed *in vitro* showed directly the oncogenic function of the RECK-B non-canonical isoform, whose predominant expression in glioma cell lines promoted anchorage-independent cell growth.

This thesis project aims at analysing the contribution of alternative splicing regulation to the definition of glioma grades 2 to 4. It will have as a focus the glioblastoma (GBM) and low-grade glioma (LGG) RNA-seq data sets from the TCGA portal.

Three main subjects will be approached. Firstly, an evaluation about whether a signature of alternative splicing that is exclusive of this layer of mRNA processing exists or rather if it is associated with the already defined glioma subtypes will be made.

Then, the prognostic value of alternative splicing in glioma will be assessed and, specifically, genes and events of alternative splicing that make good glioma prognostic markers will be identified, particularly in terms of adding prognostic value to the already known glioma clinical and molecular risk factors.

Finally, an attempt to identify potential mechanisms of alternative splicing regulation in *trans,* underlying the patterns of alternative splicing in the different samples analysed, will be made.

2 METHODS

2.1 DATA SETS

Clinical (patient- and biospecimen-related) and transcriptomic data used in this manuscript were retrieved from the data portals of TCGA ("The Cancer Genome Atlas - Data Portal," 2016) and GTEx ("GTEx Portal," 2016).

Patient and Biospecimen data files from the two glioma TCGA cohorts – Glioblastoma multiforme (GBM) and Low Grade Glioma (LGG) – included patient clinical registry information, as well as follow--up data, and detailed information about sample quality, processing and full code nomenclature to enable correct assignment of gene expression data files to tumour case. As dictated by the sample management protocols of the project, all samples entering the cohorts contained at least 70 % tumour nuclei and not more than 50 % necrosis. Samples were subjected to histological classification, including WHO grade. A description of the nomenclature used to refer to the different histological types throughout this work is made in Table 2.1.

Table 2.1 – Nomenclature code used for the different sample types of diffuse gliomas of the GBM and LGG TCGA cohorts.

Official Designation	***This manuscript's Designation***	***Acronym***	***Histology ICH id****	***WHO grade***
Diffuse Astrocytoma	Low-grade glioma grade 2	LGG2	9400/3	II
Anaplastic astrocytoma	Low-grade glioma grade 3	LGG3	9401/3	III
Glioblastoma	Glioblastoma multiforme	GBM	9440/3	IV
Diffuse Oligodendroglioma	Low-grade glioma grade 2	LGG2	9450/3	II
Anaplastic oligodendroglioma	Low-grade glioma grade 3	LGG3	9451/3	III
Oligoastrocytoma	Low-grade glioma grade 3	LGG3	9382/3	II/III

* Morphology code coming from the International Classification of Diseases for Oncology.

In turn, a summary of clinical and molecular characteristics of glioma samples is shown in Table 2.2.

RNA-seq expression data of level 3 from the same glioma cohorts referred above, processed and released by TCGA version 2 pipeline ("RNASeq Version 2 - TCGA - National Cancer Institute - Confluence Wiki," 2016), was acquired on 14/09/2015. That pipeline includes mapping of sequencing reads to the TCGA, UCSC-nomenclature based, annotation (https://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf) of the hg19 human genome assembly, using MapSplice (K. Wang et al., 2010) and quantification of expression using RSEM (B. Li & Dewey, 2011). Tables with raw gene and isoform counts, as well as normalized TPM (Transcripts per million ((B. Li, Ruotti, Stewart, Thomson, & Dewey, 2010)) for genes and isoforms, were used in different analyses.

Data associated with GTEx tissue donor subjects and biospecimens consisted of sample tissue identity and subject and sample tissue identifiers. Samples used were a total of 8555, relative to 31 different tissues from the human body, coming from 573 donors (Lonsdale et al., 2013).

Similar to what was previously described for TCGA transcriptomics data acquisition, RNA-sequencing expression data from the GTEx project was obtained as tables with raw and normalized (RPKM) read counts for genes and isoforms, on 06/01/2016 (project's data release version 6). The pipeline used for RNA-sequencing analysis included sequencing read mapping to a modified version of Gencode v12 annotation of the hg19 genome assembly: http://www.broadinstitute.org/cancer/cga/tools/rnaseqc/examples/gencode.v12.annotation.patched_contigs.gtf.gz using TopHat (Trapnell, Pachter, & Salzberg, 2009) and quantification of known transcripts through the Flux Capacitor method ("GTEx Quantifications - Flux Capacitor - Confluence," 2016; Montgomery et al., 2010).

Table 2.2 – Clinical and Molecular Characteristics of the TCGA Sample Set.

Features	*Total Cases (N = 674)*	*Publication*
Cohort	674 (700 samples)	
LGG	514	(Suzuki et al., 2015)
GBM	160	(Brennan et al., 2013)
WHO Grade		
II	250	
III	264	
IV	160	
DNA methylation Cluster Subtype		
LGm1	52	
LGm2	253	
LGm3	123	
LGm4	68	
LGm5	104	
LGm6	40	
Unknown	34	
Primary-Recurrence Availability		
Sample Pair Available	20	
Sample Pair Not Available	654	

2.2 ANALYSIS OF ALTERNATIVE SPLICING DATA

2.2.1 PSI data matrix generation

Percent splicing-index estimates were calculated with SUPPA (Alamancos, Pagès, Trincado, Bellora, & Eyras, 2014) for alternative splicing events of the SE, MX, RI, A3, A5, AF and AL types by performing the ratio of the sum of the levels of a gene's isoforms that include the regulated exon (or intron) over the sum of the levels of all the isoforms from the same gene. The alternative splicing events considered are generated by the program from the genome annotation provided and accounting for all splicing possibilities among the event types referred above. The definition of the regulated exon is done according to particular rules, specified in the software's paper.

A table of transcript isoforms quantified in TPM for the 700 GBM and LGG RNA-seq samples was assembled from individual patient files and used as input together with a TCGA genome annotation

gtf file. The two command lines of the software, *generateEvents* and *psiPerEvent,* were run with default parameters, as described in *https://bitbucket.org/regulatorygenomicsupf/suppa/src*.

2.2.2 Preparation of working PSI matrices

Resulting PSI tables were filtered for missing values, in order to eliminate very rare alternative splicing events and samples with generalized poorer quality of PSI quantification. A first filter removed alternative splicing events missing PSI calculations for more than 80 % of the samples, while a second applied filter excluded samples with missing values for more than 40 % of the AS events. Dimensions of PSI matrices for individual event types, as well as for a merged matrix containing all event types after missing values-filtering, are shown in Table 2.3.

As a final step of PSI matrix preparation, duplicated samples pertaining to the same patient were removed (with the criterion of keeping as far as possible primary tumour samples only, rather than tumour recurrences), which resulted in a final PSI table containing 17151 alternative splicing events and 659 samples.

Table 2.3 – Dimensions of PSI tables after filtering.

	Alternative Splicing Event Type							
	SE	MX	RI	A3	A5	AF	AL	All events
Events (N)	10700	118	713	2093	1740	1553	234	17151
Samples (N)	694	693	698	693	699	694	700	686

2.2.3 Differential alternative splicing analysis

Analysis of differential alternative splicing regulation across LGm subtypes was performed using the non-parametric Kruskal-Wallis statistical test for the equality of medians. This test can be applied when distributions of the variable under study are not normal, which is the case for PSI values (Rosner, 2011).

A PSI matrix containing all 17151 alternative splicing events quantified for the 627 samples with known, LGm group was used (Ceccarelli et al., 2016). The Kruskal-Wallis test was applied through function kruskal.test*,* the Kruskal-Wallis test implementation from the R package *stats*, and, each alternative splicing event has been tested using a list of six PSI vectors, one per LGm group. Adjustment for multiple hypotheses testing was performed by the Benjamini & Hochberg False Discovery Rate (FDR) correction (Benjamini & Hochberg, 1995), using the function *p.adjust* from the R package *base,* and both Kruskal-Wallis test statistic and FDR values were kept for downstream analyses.

For the selection of alternative splicing events showing a minimum PSI median difference between groups of 0.1, all 15 combinations of median differences between the six LGm groups were calculated for each of the 17151 alternative splicing events and, finally, events were selected if they had any of these differences reaching an absolute value of 0.1.

2.3 ANALYSIS OF GENE EXPRESSION DATA

Exploratory and differential gene expression analyses were performed using functions from Bioconductor packages edgeR ("edgeR: a Bioconductor package for differential expression analysis

of digital gene expression data," 2016), which was created specifically for assessment of differential expression from count data such as RNA-seq, and limma (Ritchie et al., 2015).

2.3.1 Preparation of working gene expression matrices

A gene expression matrix was assembled from the 659 individual sample files containing gene raw read counts for the 20531 genes included in the TCGA genome annotation.

For performing exploratory analysis, the referred matrix was used together with a vector of tumour grade identifiers and a matrix of gene identifiers to create an edgeR-specific DGEList object type. Then, a filter for lowly expressed genes was applied: genes having more than 1 cpm (counts per million) in at least 160 samples, the sample size of the smallest group considered (the GBM samples), were kept in the DGEList, a criterion suggested in the edgeR user guide with the rationale that it guarantees that any gene in the analysis can only be considered differentially expressed between groups if consistently detected with a good signal in all samples from at least one group. The resulting matrix had 15189 genes. Read count normalization was carried out running the command *calcNormFactors*, with default settings, which performs a normalization of count data, taking into consideration samples library sizes and library compositional differences. Sample compositional differences are accounted for through the application of the trimmed mean of M-values (TMM) method (Robinson & Oshlack, 2010), which estimates scaling factors for the library sizes (and thus for the total RNA output of each sample) that will minimize the log-fold changes between samples for most genes.

edgeR uses negative binomial distributions to model the read counts from each gene in a library, the expected counts of the distribution being given by the parameter probability of success of finding the gene in the whole library multiplied by the library size. The biological coefficient of variation across samples for that expected count is given by the square root of the dispersion parameter of the negative binomial distribution. Dispersion estimates for each tag (gene) were calculated with the function *estimateDisp* and a design matrix built for the factor being considered: tumour grade. Specifically, using the command *model.matrix*, each sample took the value 1 for the level 2,3, or 4 to which it belonged to and zero otherwise. A detailed user's guide for performing RNA-seq analysis using EdgeR can be consulted at the package's page of Bioconductor website ("edgeR," 2016).

For differential gene expression analysis across LGm subtypes, the initially prepared table of raw RNA-seq read counts for the glioma samples, this time including only the 627 samples of known LGm subtype, was used together with a vector of sample LGm group identifiers and a matrix of gene identifiers to create a DGEList object. Then, a filter for lowly expressed genes was applied: genes having more than 1 cpm in at least 40 samples, the sample size of the smaller group under study, were kept in the DGEList. The resulting matrix had 15957 genes. Read count normalization through the trimmed mean of M-values (TMM) method (*see* explanation at the beginning of this section) was carried out running the command *calcNormFactors*, with default settings.

Dispersion estimates were calculated and a design matrix built for the factor in study: LGm affiliation, following the same procedures already described.

2.3.2 Differential gene expression analysis

Given the DGEList object (containing (1) a raw count table, (2) a gene list, (3) sample normalization factors and (4) gene-specific dispersion estimates) and the design matrix, negative binomial generalized linear models were fitted to the expression estimates for each gene, using function

glmQLFit. Then, a one-way ANOVA-like empirical Bayes quasi-likelihood F-test to detect genes differentially expressed between any of the subtypes was carried out, using the *glmQLFTest* function over the output models and a contrast matrix consisting of the five pairwise comparisons between LGm groups 1,2,4,5,6 and LGm3. To get a summary table of the results, the *topTags* function was used.

Differential gene expression was considered for genes whose F-test for the equality of expression between all LGm levels returned an adjusted p-value below 0.01 and that in addition had a log2-fold change of at least 1 in relation to the least malignant LGm subtype: LGm3.

2.4 EXPLORATORY DATA ANALYSIS

2.4.1 Alternative splicing *vs* Gene expression correlation analysis

The strength of dependence of splicing ratios on the levels of transcriptional output was assessed using rank-based two-sided Spearman correlation tests for each vector pair of PSIs and corresponding gene expression levels (in counts per million or cpm). These tests were run using function *cor.test* from stats CRAN R package. A visual inspection of the relation between PSIs and cpms for selected events was made using the scatter plot function *smoothScatter* from the CRAN R package graphics.

2.4.2 PSI variances

Variances of quantification of each alternative splicing event within selected groups of samples were calculated with function *var* from CRAN R package stats and their distributions were visualized using functions *densityplot* and *bwplot* from CRAN R package lattice.

2.4.3 Principal Component Analysis

Principal Component Analysis of PSI matrices was performed using function *PCA* from the CRAN R package FactoMineR. For each alternative splicing event, PSI values were first subjected to centering around zero, using function *stdize* from the CRAN package pls. The *PCA* function was then run, without scaling, on the resulting matrix.

The same functions were used for Principal Component Analysis of gene expression matrices. A table of read counts was exported in counts per million (cpm) from the respective edgeR-specific DGEList object and further log2-transformed using the *voom* function from R package limma. For each gene, expression levels in a logarithmic scale were then centered around zero and also not subsequently scaled when running *PCA*.

Scores (samples) and loadings (events/genes) for each principal component were extracted from the "coord" matrix of "var" list and "ind" list, respectively.

Sample scores for each principal component of interest were plotted using functions from the ggplot2 and gridExtra CRAN R packages.

2.5 FUNCTIONAL ENRICHMENT ANALYSIS

In order to identify cellular pathways or biological processes strongly associated with DNA-methylation subtype distinction or degrees of tumour malignancy, gene set enrichment analysis (GSEA) was used (Subramanian et al., 2005). This functional analysis *in silico* method is based on the expectation that, given a set of genes *S* belonging to a common functional category and a large list of genes *L* ranked according to the strength of association with the distinction of classes under study (e.g. according to statistical significance of differential expression between classes), the functional category represented by *S* will be relevant for this distinction if its genes accumulate in a biased, non-uniformly distributed, way at either of the extremities of the ranked *L* list. The measure of the relevance is given by an enrichment score (ES) that is the maximum absolute value attained by a running-sum statistic obtained along a random walk through the ranked *L* list, in which the statistic is incremented in steps involving genes from the gene set *S* in study and decremented otherwise. The significance of the test is assessed from a null distribution of the ES obtained by performing sample or gene permutations followed by ES calculations.

GSEA was carried out using the GSEA-P application, kept by Broad Institute, using gene sets from the MSigDB database version 5.1, also maintained by the same institution, namely for KEGG pathways, Reactome pathways and Gene Ontology Biological Process terms. The pre-ranked analysis mode was run, without weighted steps for ES calculations and a minimum number of 15 genes from the gene set under test having to be present in the list under study. The metrics used for the analysis of functional category enrichment among genes/alternative splicing events that better differentiate LGm subtypes were the F-statistic and Kruskal-Wallis rank-test statistic as obtained from differential gene expression and differential splicing analysis, respectively. For the differentiation of levels of malignancy, the metrics used was principal component loadings for genes/events of alternative splicing, the module of the loadings having been used in the latter case. The enrichment of a gene set was considered significant when the associated FDR adjusted p-value was below 0.05.

2.6 SUPERVISED SAMPLE CLASSIFICATION

The PAM algorithm implemented in the pamr R package (Tibshirani, Hastie, Narasimhan, & Chu, 2002) was run on the 17097 PSIs of non-zero variance across the 627 samples with assigned DNA-methylation cluster subtype (LGm group). This supervised learning method relies on an approach called the nearest shrunken centroid classification. Nearest shrunken centroid classifiers function by computing, for each variable and for each class, a centroid which consists of a coefficient of variation (mean/standard deviation). Then when a sample class has to be predicted, the Euclidean distance from the values it takes for each variable to the corresponding centroids of each class is calculated and the sample gets assigned to the closer class. The PAM algorithm includes an extra feature, which is a shrinkage procedure, which consists of picking a number of values and subtracting them one at a time from each class centroid. At each step the classifier will be revaluated by cross-validation, to check for the shrinkage value that generates less prediction error. Some genes are eliminated from the classifier due to shrinkage and, since the lowest error level classifier typically has a non-zero shrinkage, it will consist of a subset of the initial variables submitted to build it. The steps of this analysis included a training step, run through the command *pamr.train,* a cross-validation step, run by the function *pamr.cv*, and a final FDR estimation for all the classifiers at multiples shrinkages, using *pamr.fdr*. Finally, the classifiers chosen were in each case (for each training set used) the one producing fewer errors on cross-validation and its list of classifying genes compiled using the *pamr.listgenes* function.

2.7 SURVIVAL ANALYSIS

2.7.1 Kaplan-Meier curves

Kaplan-Meier survival curves are built through calculation of survival functions, i.e. probabilities of survival up to a given time. These survival values are usually obtained using the Kaplan-Meier estimator, which consists of, for each time t, multiplying the conditional probability of surviving up to time t given that one survived until time t-1 by the survival value at time t-1 (Rosner, 2011). In these calculations, the conditional probabilities for survival at each time are obtained excluding patients that were censored during the last time interval between collections of patients' follow-up data. Survival curves were created by plotting values of the survival function for the different strata, obtained with the help of CRAN R package survival. First, a *Surv* object was created with patient right-censored overall survival data. This object was then passed to the *survfit* function using a categorical factor vector as the formula to specify the different patient strata. Survival curves were plotted using standard functions.

2.7.2 Cox regression models

With the aim of finding the relationship between patient's overall survival and exposure to certain factors, namely increasing levels of gene expression or PSI values for an alternative splicing event, Cox proportional-hazards models were used (Prentice, 1992). These models allow to estimate the ratio between the hazards, i.e. the instantaneous probability of an event (e.g. a death event) at time t+Δ, given survival until time t, of two subjects that differ by one unit of exposure to a potentially impactful variable on survival. The Cox proportional-hazards models are regression models with the following general formula, given a set of explanatory variables k:

$$h(t) = h_0(t)\exp(\beta_1 x \, x_1 + \ldots + \beta_i x \, x_i + \ldots + \beta_k x \, x_k),$$

where $h_0(t)$ and $h(t)$ are the hazards at time t of having, respectively, a baseline value for the k independent explanatory variable(s) and a baseline value incremented of x_i units for the same variable(s). The coefficient β_i represents the hazards ratio for each particular explanatory variable or risk factor and is assumed to remain constant throughout time in order for the application of these models to be reliable. The null hypothesis that each β_i is equal to zero *vs* the alternative hypothesis that it is different from zero can be tested using the test statistic $Z = \beta_i / se(\beta_i)$ and conducting a two-sided significance level α test. Confidence intervals for the β_i estimation can be given by (e^{c1}, e^{c2}), where $c1 = \beta_i - z_{1-\alpha/2} se(\beta_i)$ and $c2 = \beta_i + z_{1-\alpha/2} se(\beta_i)$, with $z_{1-\alpha/2}$ corresponding to the quantile of probability $1-\alpha/2$ of a standard normal distribution (Rosner, 2011).

Cox models to study the value of various independent variables on patient's overall survival were derived using function *coxph* from the *survival* package on a Surv object, built as specified above. A description of the main generated models is presented in Table 2.4.

The levels of nominal risk factors for each patient were specified using categorical, factor class, vectors with levels designations, while the levels of continuous risk factors were specified as numeric vectors. Principal component sample scores, as well as PSI values, were used directly for Cox-model derivation. Gene expression levels were used in logarithmic scale, a procedure shown to perform correctly in a paper that compared several transformation and scaling methods for application to RNA-seq data for Cox model creation purposes (Zwiener, Frisch, & Binder, 2014).

Table 2.4 – Description of the main Cox proportional-hazards models derived.

Variable(s)	*Model equation*
Gene expression PC1 sample score	$h(t) = h_0(t)\exp(\beta_{PC1} \times \text{PC1 sample score})$
Gene expression PC2 sample score	$h(t) = h_0(t)\exp(\beta_{PC2} \times \text{PC2 sample score})$
WHO grade	$h(t) = h_0(t)\exp(\beta_{Grade} \times \text{Grade})$
Gene expression level	$h(t) = h_0(t)\exp(\beta_{Gene} \times \text{GElevel})$
Percent spliced-in ratio (PSI)	$h(t) = h_0(t)\exp(\beta_{ASevent} \times \text{PSI})$
Percent spliced-in ratio and Gene expression level	$h(t) = h_0(t)\exp(\beta_{Gene} \times \text{GElevel} + \beta_{ASevent} \times \text{PSI})$
Age at diagnosis	$h(t) = h_0(t)\exp(\beta_{Age} \times \text{Age})$
DNA-methylation cluster	$h(t) = h_0(t)\exp(\beta_{DNAmetclust} \times \text{DNA\_met\_clust})$
Percent spliced-in ratio, Gene expression level, DNA-methylation cluster, Grade and Age	$h(t) = h_0(t)\exp(\beta_{Gene} \times \text{GElevel} + \beta_{ASevent} \times \text{PSI} + \beta_{DNAmetclust} \times \text{DNA\_met\_clust} + \beta_{Grade} \times \text{Grade} + \beta_{Age} \times \text{Age})$
Gene expression level, Grade and Age	$h(t) = h_0(t)\exp(\beta_{Gene} \times \text{GElevel} + \beta_{Grade} \times \text{Grade} + \beta_{Age} \times \text{Age})$
Gene expression level, Percent spliced-in ratio, Grade and Age	$h(t) = h_0(t)\exp(\beta_{Gene} \times \text{GElevel} + \beta_{ASevent} \times \text{PSI} + \beta_{Grade} \times \text{Grade} + \beta_{Age} \times \text{Age})$

2.7.3 Venn diagrams

Venn diagrams were produced using a specific tool for that purpose, produced by the Girke Lab ("Graphics and Data Visualization in R," 2016), whose R script is available at the URL http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/My_R_Scripts/overLapper.R. This script requires as input a list of vectors corresponding to the sets to enter the diagram and diagrams are produced using the *vennDiagram* function from the limma package.

2.8 STUDY OF ALTERNATIVE SPLICING REGULATION IN *TRANS*

2.8.1 Correlations between RBP gene expression and exon inclusion levels

Analyses of the correlation between alternative splicing event PSIs and RBP gene expression levels were performed as described in 2.4.1, except that the gene expression measurement unit used was TPM. Only samples having a minimum RBP gene expression level of 1 TPM were used. The concordance between correlation test results for the GTEx and TCGA datasets was made through plotting the logarithm of the correlation test FDR adjusted p-values for the alternative splicing events common to the two datasets, with log-FDR values relative to negative correlations being kept negative and those relative to positive correlations taking a plus sign.

2.8.2 Mapping of RBP binding motifs along the genome using FIMO

Binding motifs for 224 RBPs have been identified in this work (Ray et al., 2013), each one having been represented as a 7-mer (7 nucleotides long) position-specific scoring matrix (PSSM), which is a way of defining biological patterns that allows for different levels of nucleotide degeneration across the positions defining the motif. As such, for a seven nucleotides long motif, a PSSM matrix has four rows representing each nucleotide and seven columns representing each position of the motif, while the values of the matrix are probability scores that determine the frequencies at which a nucleotide appears at each position of this particular motif. In the particular case of the RNA binding motifs described in this paper, these sequences were identified by protein-RNA competition assays, and so

PSSM scores reflect binding preferences for each RBP. These PSSM for each RBP of interest were used by the FIMO tool from the MEME suite for motif analysis to map motif occurrences along the human genome. The *fimo* command was run using as input files motif PSSMs in MEME format and the hg19 genome sequence in FASTA format, and default options, except for a p-value threshold of $1x10^{-3}$.

2.8.3 Quantification of putative alternative splicing event targets for different RBPs

Identification of putative target alternative splicing events for the RBPs of interest was done through selecting those containing binding motifs identified by FIMO at a $p=1x10^{-3}$ threshold and having their splicing ratios correlated with RBP gene expression in both glioma TCGA and multitissue GTEx datasets at an FDR cut-off of 0.01.

2.8.4 Definition of regulatory regions for RNA splicing map generation

Eight regulatory regions were considered for the general exon-skipping alternative splicing event. Relevant features taken into account were the alternative/regulated exon, its two flanking introns and the upstream and downstream constitutive exons (Figure 2.1). Given these, segments of 150 bp of intronic sequence flanking the three exons as well as of 50 bp of exonic sequence spanning the beginning and end of these exons were considered. These lengths implied a minimum intron length of 300 bp and a minimum exon length of 100 bp in order for regulatory sequences not to overlap. The length of the segments was chosen based on literature reviewing to find the most usual relevant positions described in RNA splicing maps. However, a compromise was made between using genomic segments long enough to likely contain splicing regulatory sequences and short enough to avoid excluding too many events from the analyses due to exon or intron length limitations.

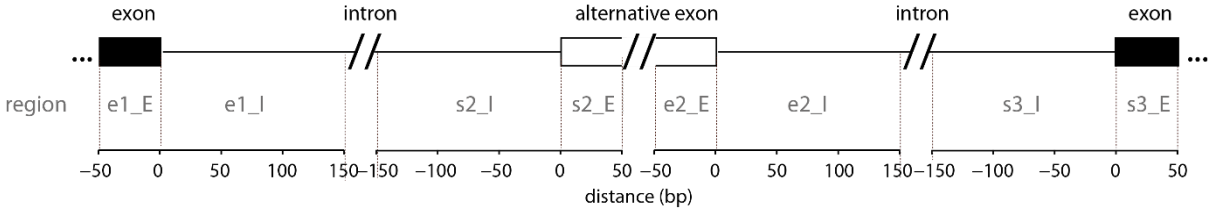


Figure 2.1 – Definition of regulatory regions for a general event of exon-skipping (SE).e1_E – exonic region corresponding to last 50 nucleotides of first constitutive exon; e1_I/s2_I – intronic region corresponding to first/last 150 nucleotides of intron located upstream from alternative exon; s2_E/e2_E – exonic region corresponding to first/last 50 nucleotides of alternative exon, e2_I/s3_I – intronic region corresponding to first/last 150 nucleotides of intron located downstream from alternative exon; s3_E – exonic region corresponding to first 50 nucleotides of second constitutive exon

2.8.5 Determination of the best correlation test and motif binding threshold parameters for generating each RNA splicing map

Identification of likely targets of RBP action was based on the detection of events of alternative splicing whose PSIs correlated with RBP abundance and that effectively contained RBP binding motifs in their regulatory regions.

To select the group of significantly correlated alternative splicing events, an FDR adjusted p-value threshold for the correlation between PSIs and RBP gene expression across samples was used. To define the group of RBP binding motifs, a p-value cut-off for the stringency of motif identification by FIMO was in turn used. Specifically for this latter, lower p-value thresholds were used for less degenerate and thereby less ambiguously detectable 7-mers, while higher p-value thresholds allowed the detection of 7-mer motifs defined by PSSM models reflecting looser combinations of

nucleotides. These two p-value parameters were allowed to change until maximum signal for RBP regulation was reached, as described below.

Once having set these two parameters, a Fisher's exact test for each of the eight regulatory regions was run, having as null hypothesis that alternative splicing events whose PSIs correlate with the expression of an RBP are independent of them containing in that regulatory region a binding motif for the same RBP. One-sided tests were performed to explore the alternative hypothesis that the proportion of alternative splicing events correlated with RBP expression is higher when binding motifs are present than when not. To distinguish between putative enhancing and silencing roles of the RBP binding to the regulatory region in the inclusion of the exon, tests for positively and negatively correlated alternative splicing events were performed separately, always having the group of events that did not significantly correlate with RBP expression as the null sample. To identify the alternative splicing events that contained at least one RBP binding motif for each regulatory region, firstly, genomic coordinates of both binding motifs mapped by FIMO and regulatory regions for the annotated splicing events were used to create "genomic range" R objects, and secondly, the overlap between the two features was checked. The representation of genomic ranges was implemented using function *GRanges* and range overlaps were obtained with function *findOverlaps,* both from the Bioconductor package GenomicRanges.

Maximizing the significance of the Fisher's tests referred above was achieved by trying series of different correlation FDR adjusted p-values and RBP binding motif p-values. These series corresponded to 5 % quantile steps of correlation FDR values and the 5, 20, 35, 50, 65, 80 and 95 % quantiles of FIMO p-values for the binding motif under consideration. To visually search for ranges of parameters that jointly and consistently appear to maximize the significance of the tested association, heat maps were drawn for each set of tests applied to each regulatory region for both positive and negative correlations, using function *heatmap.2* from CRAN R package gplots. Two sets of parameters were selected to use in the generation of RNA splicing maps: the two that produced a lower p-value for each of two regulatory regions. This selection from two rather than one region was found useful to ascertain the robustness of the inferred map.

Finally, using a single combination of correlation FDR adjusted p-value and motif p-value selected in the previous step, RNA splicing maps were based on the p-values of Fisher's exact tests applied to each of the 800 nucleotide positions in the general alternative splicing event defined by the concatenation of the regulatory regions. Tests were made on 50 nucleotide-spanning sliding windows, centered in the position of interest. The RNA splicing maps were plotted with graphing functions from the ggplot2 package.

3 RESULTS

3.1 SIGNATURES OF ALTERNATIVE SPLICING IN GLIOMA

This section will be dedicated to characterizing the overall patterns of splicing regulation in the studied glioma cohort. Departing from PSI matrices for a large set of alternative splicing events, exploratory analysis, namely using multivariate methods, will be performed, looking at how this data organize in relation to other clinical variables and molecular classifications, with particular focus on the comparison of PSI data with gene expression data. Subsequently, in order to identify the alternative splicing events that vary according to the pan-glioma DNA-methylation cluster subtype, along with potential *trans*-activators of these events, differential splicing and gene expression analysis between groups will be carried out and interesting new findings outlined.

Data tables with alternative splicing event PSI values and gene expression levels used throughout this section represent 659 glioma patients, or 627 in analysis where DNA-methylation subtypes must be specified, as this classification was not available for the remaining 32. As for the alternative splicing events included in the analysis, these are a total of 17151, representing 7349 regulated cognate genes, of the types skipped exon, mutually exclusive exons, retained intron, alternative 3' splice site, alternative 5' splice site, alternative first exon and alternative last exon, as defined in the manuscript's introduction and whose individual numbers are shown in the Methods section. Alternative first exon events, despite having a less clear and for sure weaker association with splicing machinery function, were nevertheless included in the present study given their potentially equally important contribution to gliomagenesis as the one given by the other event types. Gene expression data tables used contained 15189 or 15957 genes (*see* Methods).

3.1.1 Determination of the level of dependence of alternative splicing on the expression of cognate genes

Transcript levels for a given gene are good indicators of the overall abundance of transcripts that serve as the substrate for splicing machinery action, and can also work as an indirect indication of the rates of RNA synthesis, known to be relevant for splice site recognition by the spliceosome, namely when considering regulated splicing for which there is splice site competition.

Since the main focus of this thesis is to evaluate the contribution of alternative splicing regulation in glioma, an identification of alternative splicing events clearly independent of gene expression was carried out. In order to determine the potential strength at which transcription influences alternative splicing, or the reverse, two-sided Spearman correlation tests between each splicing event PSIs and the levels of its corresponding gene transcripts were made.

A big portion of the events had their PSIs correlated in a consistent way with their gene expression, 11262 (66 %) and 9784 (57 %) out of 17151 having showed a significant Spearman correlation at an FDR below 0.05 and 0.01, respectively. However, as can be observed in the plot from Figure 3.1, significant events had correlation coefficients predominantly very low: minimum value of 0.09 at FDR < 0.05 and of 0.11 at FDR < 0.01, suggesting there is no strong mutual influence between transcriptional activity and splicing regulation in glioma samples. Considering an FDR cut-off of 0.05, there were still 5837 alternative splicing events (representing 3761 genes) whose regulation was not significantly correlated with their own gene expression. In terms of the sense of the association between gene expression and alternative splicing ratios, 55 % of the events had higher PSIs with

increased gene expression, and there were also more events showing positive correlation with gene expression among the ones with rho Spearman coefficients with an absolute value above 0.5 (63 %).

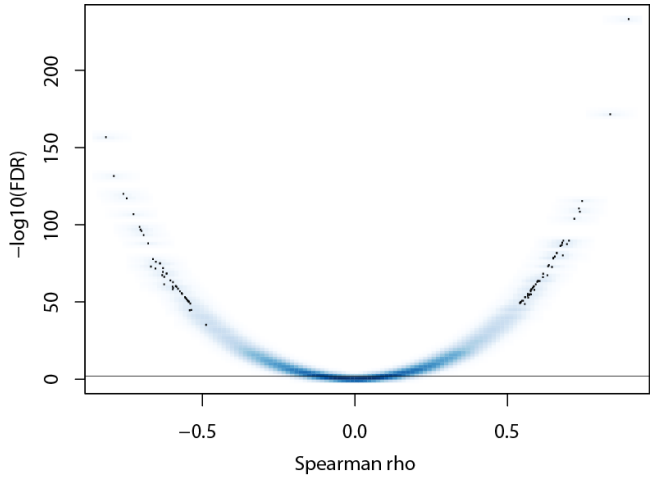


Figure 3.1 – Correlation between PSIs of AS events and levels of gene expression of cognate genes Scatter plot of FDR of Spearman correlation tests vs correlation coefficient is shown.

Some of the top significant correlations corresponded to alternative first exon events, which are known to be highly linked to transcription initiation, rather than with splicing regulation. Events of these types account for only ~8 % of all the alternative splicing events studied but could be thought of as being a major contributor to the high proportion of statistically significant alternative splicing event-cognate gene expression correlations found. In order to understand if this was the case, quantification of significant correlations was redone this time including all event types other than alternative first and last exon types. Again, not only the majority of alternative splicing events showed a consistent correlation between their PSIs and gene expression, but the proportions of significant hits were exactly the same as detected when considering all event types: 10071 (66 %) and 8737 (57 %) out of 15364 events showed a significant Spearman correlation at an FDRs of 0.05 and 0.01, respectively.

3.1.2 Assessment of the extent of alternative splicing regulation/dysregulation in glioma

The variance of PSI values detected for a given alternative splicing event indicates how strongly this is subject to regulation in a considered group of samples, and may therefore also reflect the potential for this event to allow tumour class/subtype stratification. To understand the extent of alternative splicing regulation in glioma, and in particular among alternative splicing events whose PSIs are determined mostly by other effects rather than their own gene expression, density plots and boxplots for PSI variances were drawn (Figure 3.2).

Most alternative splicing events had PSI variances below 0.012 (standard deviation < 0.11), the 75 % quantile for all AS events, with a minority of events displaying a quite high PSI variance, going from 0.031 (Q3 + 1.5 IQR) up to 0.200 (maximum value). The distribution of variance of the group of alternative splicing events whose PSIs were not (or were weakly) correlated with the expression of their cognate gene was overall lower than the one of the group of events for which this correlation was present (Figure 3.2). This observation suggested that in the former group of events there might then exist a lower proportion to be differentially spliced between glioma subtypes and thus be able

to work as a good glioma subtype biomarker. The existence of a glioma alternative splicing signature that is independent of gene expression will be assessed below.

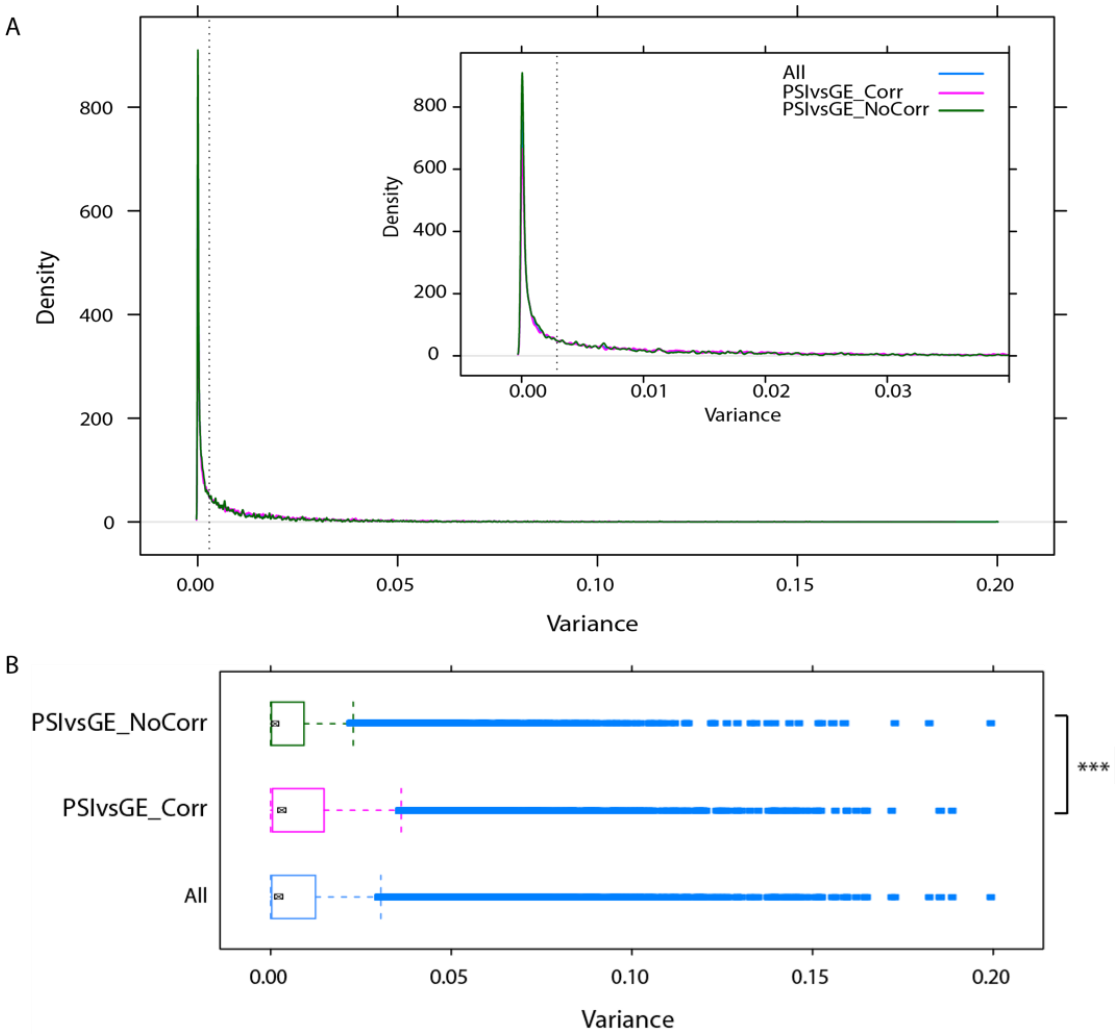


*Figure 3.2 – Variance of AS events measurements in the TCGA glioma cohort. Density plots (A) and boxplots (B) of PSI variances for all 17151 AS events considered in this study (All), for 9784 events whose PSIs correlate with gene expression (PSIvsGE_Corr) and for 7367 events whose PSIs do not correlate or correlate more weakly with GE (PSIvsGE_NoCorr). Vertical dashed line in A represents the median variance value of 0.0029. Kolmogorov-Smirnov test for the null hypothesis the equality of the two distributions was used to assess statistical significance. *** - p-value<1x10^{-3}.*

An event's PSI variance across tumour samples may also indicate the level of dysregulation of mRNA splicing in relation to normal tissue samples or, in the case of this pan-glioma study that does not include normal reference tissues, may indicate major differences in splicing machinery function between relevant tumour subtypes. For example, a glioma subtype could have the enhancing or suppressive role from an important splicing factor disrupted, or else the splicing machinery could have its efficiency affected at a particular level. This kind of alterations could potentially be detected through a change in the distribution of PSI variances in that subtype.

Two particularly interesting glioma classification systems scrutinized for associations with this kind of dysregulation were tumour grade and DNA-methylation subtype (Figure 3.3-3.4).

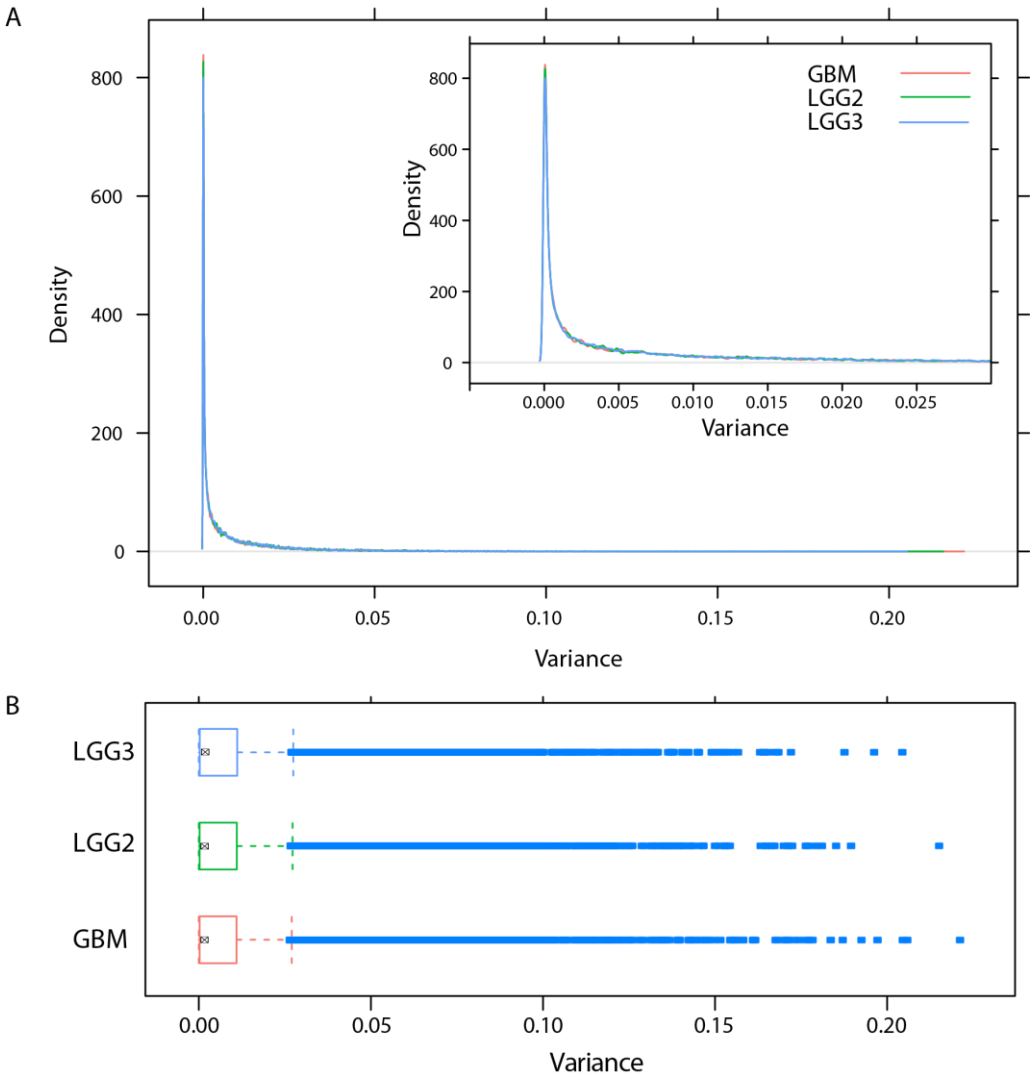


Figure 3.3 – Variance of AS events measurements in the TCGA glioma cohort. Density plots (A) and boxplots (B) of PSI variances for all alternative splicing events considered in this study, in glioblastoma multiforme (GBM), grade III low grade glioma (LGG3) and grade II low grade glioma (LGG2) samples.

Both along tumour grades and LGm groups, PSI variance distributions appeared similar overall. However, in the case of LGm subtypes some differences could be observed in the size of the low variance density peaks and the positions of the 75 % quantile. Indeed, LGm subtypes 2 and 3, together with samples of unknown LGm subtype, which were all GBM cases, had higher numbers of low alternative splicing event PSI variances, while LGm 1 and 6 subtypes displayed variance distributions predominantly with higher values (Figure 3.4). This overall behaviour showed by LGm1 and LGm6 variances could indeed be the result of a loss of function of alternative splicing regulation, but could also arise from a higher intragroup molecular heterogeneity.

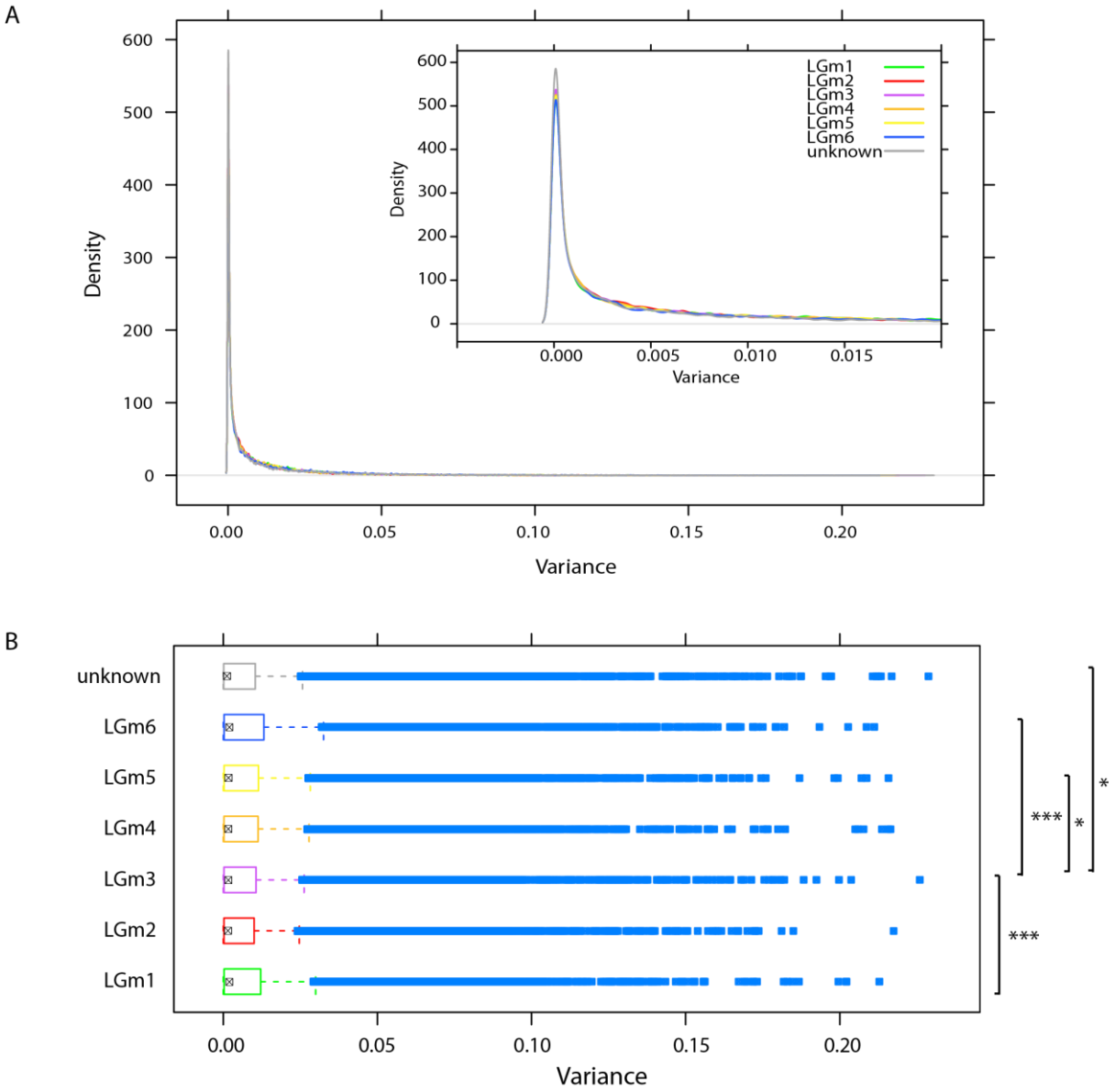


*Figure 3.4 – Variance of AS events measurements in the TCGA glioma cohort. Density plots (A) and boxplots (B) of PSI variances for all alternative splicing events considered in this study, in samples from each of the six LGm subtypes. Kolmogorov-Smirnov test for the null hypothesis the equality of the two distributions was used to assess statistical significance. * - p-value<5x10^{-2},*** - p-value<1x10^{-3}.*

3.1.3 A portrait of gene expression and alternative splicing in glioma

In order to understand how the two levels of transcriptomic data, gene expression levels and alternative splicing event PSIs, varied along glioma samples and specifically which clinical and molecular parameters contributed the most to that variation, principal component analysis (PCA) was used. This method was applied first to the full gene expression and PSI matrices, and subsequently to individual alternative splicing event types in order to identify possible differentially affected aspects of alternative splicing regulation in glioma.

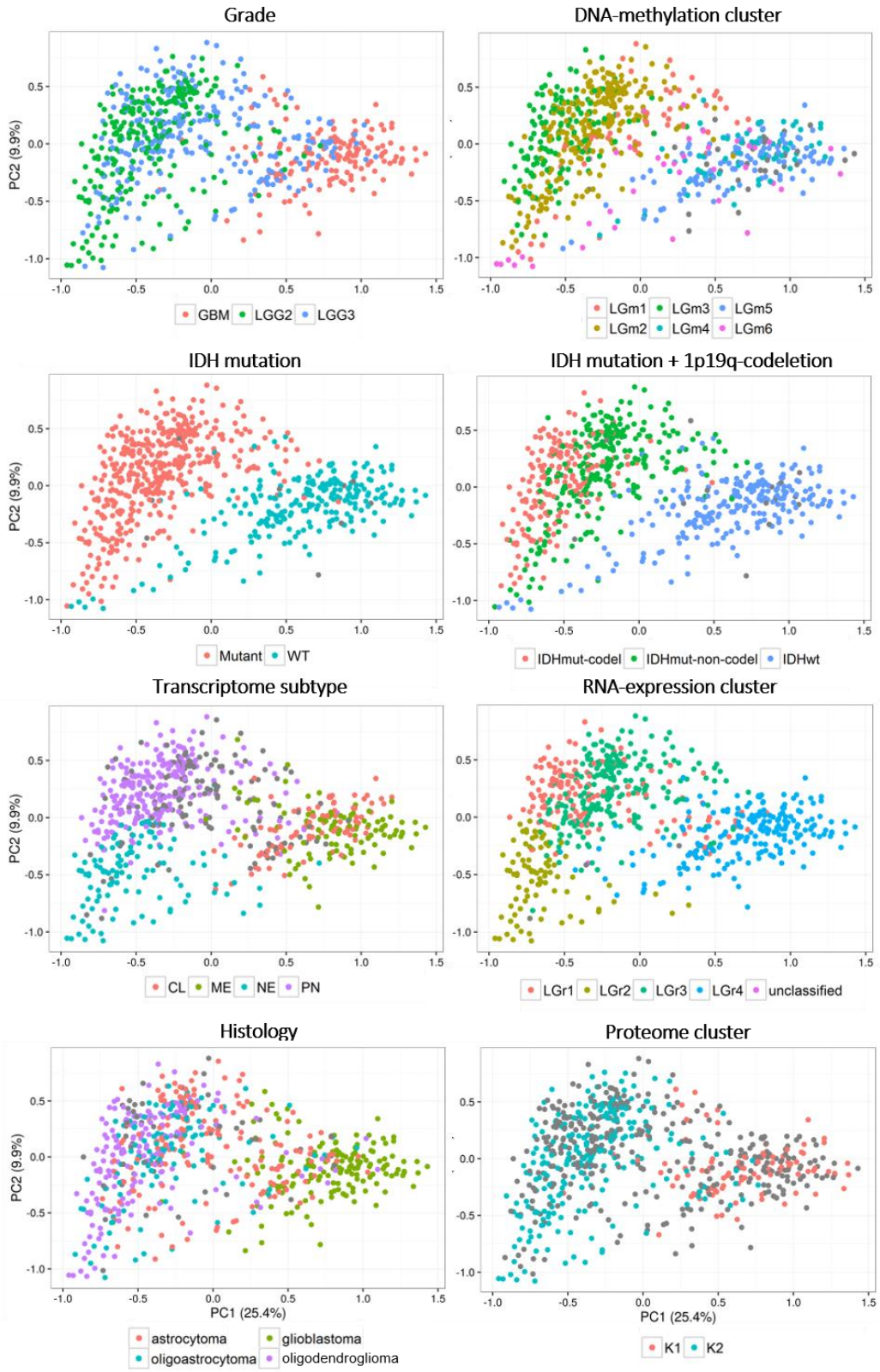


Figure 3.5 – Principal Component Analysis scatter plots of gene expression in glioma. Voom normalized, log2 counts per million of the 659 glioma cases and from 15189 non-zero variance expressed genes were analysed. Colour codes refer to eight different glioma classification systems, retrieved from TCGA metadata files and (Ceccarelli et al., 2016).

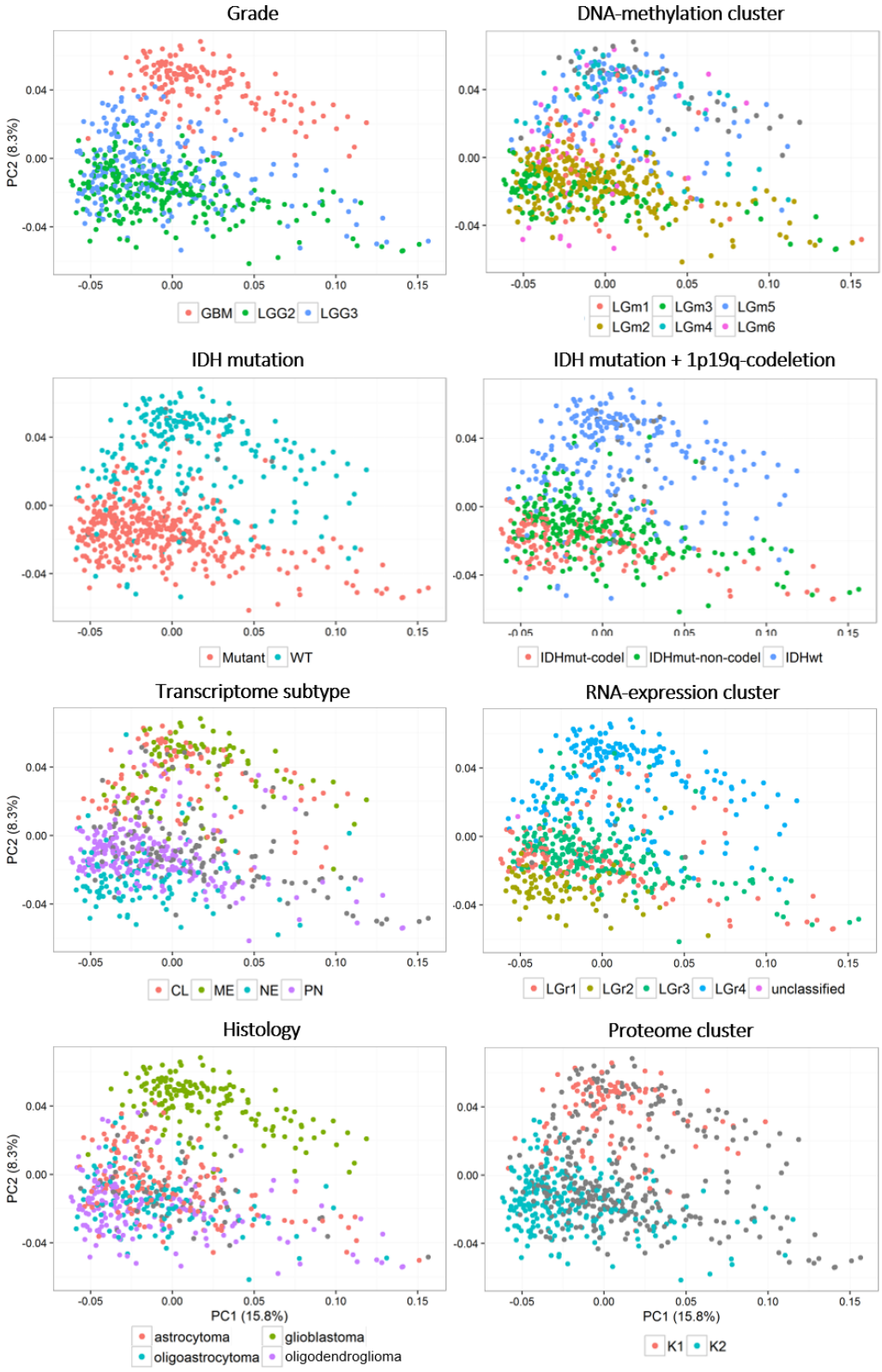


Figure 3.6 – Principal Component Analysis scatter plots of PSIs of the alternative splicing events measured in glioma. PSIs of the 659 glioma cases and from 17097 non-zero variance alternative splicing events were analysed. Colour codes refer to eight different glioma classification systems, retrieved from TCGA metadata files and (Ceccarelli et al., 2016).

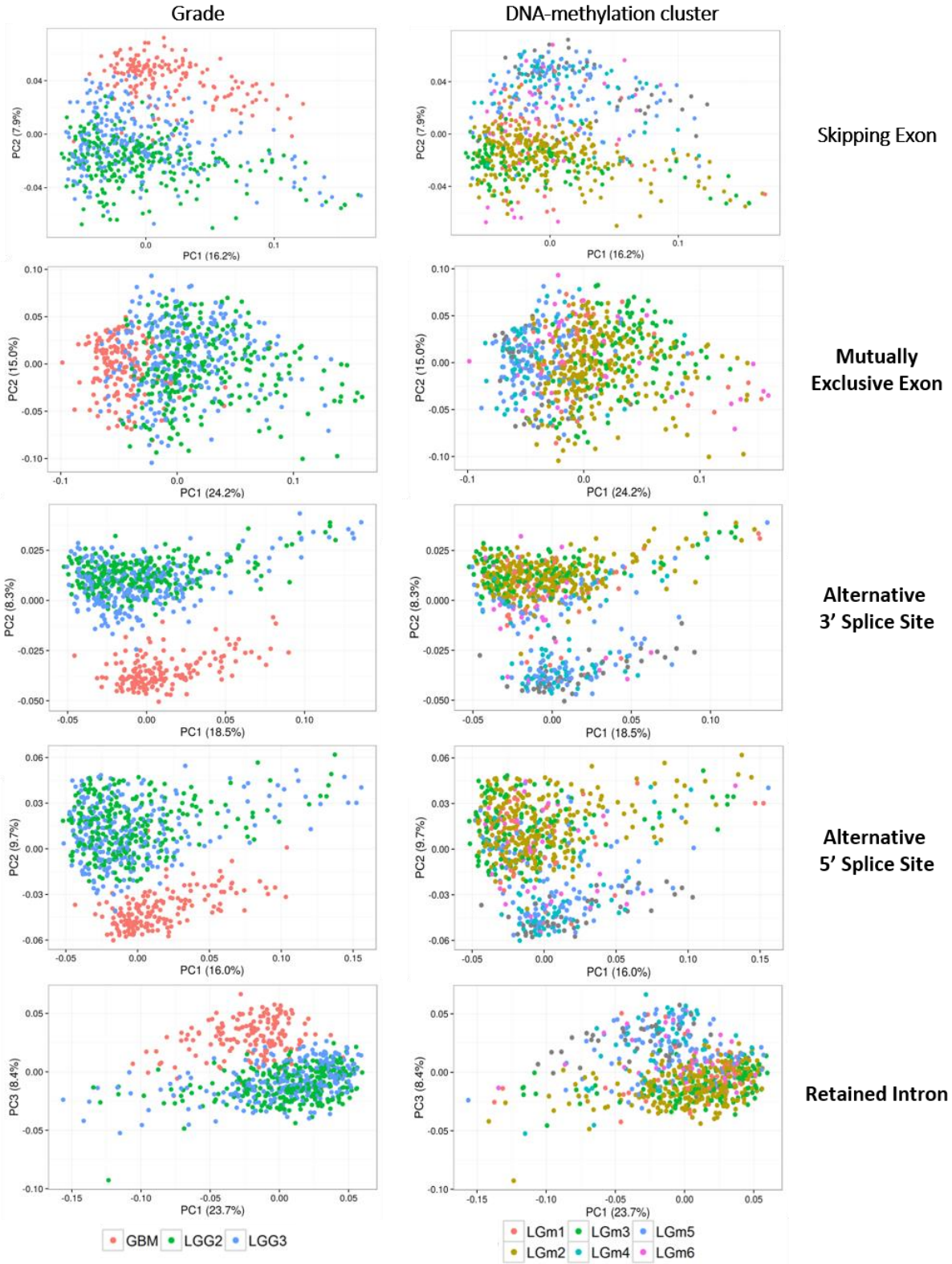


Figure 3.7 – Principal Component Analysis scatter plots of PSIs of the alternative splicing event types measured in glioma. PSIs from non-zero variance alternative splicing events were analysed. Colour codes refer to grade and DNA-methylation cluster sample assignment, retrieved from TCGA metadata files and (Ceccarelli et al., 2016).

Two-dimensional scatter plots of sample mappings along pairs of principal components explaining more than 1% of data variance each were inspected for the definition of sample clustering trends. In most cases, no distinct groups of samples were observed. Still, there was one principal component for each of the types of data looked upon that distributed samples in a quite consistent way along ordered tumour grades, from the least malignant grade 2 to the most malignant grade 4.

Selected PCA plots along the most relevant principal components, coloured according to important clinical and molecular classification systems, are shown in Figures 3.5 for gene expression, 3.6 for all alternative splicing events and 3.7 for individual alternative splicing event types.

Both gene expression and alternative splicing data are able to separate samples of WHO grade 2 (LGG2 samples) from the ones of grade 4 (GBM), along principal components 1 and 2, respectively. Samples of grade 3 (LGG3) in turn appear spread along these dimensions, which as a whole creates a gradient that goes from the less malignant to the more malignant samples. Different DNA methylation subtypes all had some level or superposition along gene expression principal component 1 and alternative splicing principal component 2. However, subtypes LGm2 and LGm3 formed a cluster that is well separated from another that included subtypes LGm4 and LGm5. Samples from subtypes LGm1 and LGm6 appear not to separate at all along the principal components shown and that have been inspected, which suggests the DNA-methylation markers used in the LGm epigenetic classifier to define these two subtypes are not strongly related with transcriptomic data. As for *IDH* mutation status, most wild type (LGm1-3) and mutant (LGm4-6) samples very evenly occupy two distinct hemi planes both in the gene expression and in the alternative splicing plots, with just minor outliers from both *IDH*-wild type and -mutated groups, which can be seen from the DNA-methylation cluster plot to correspond to samples from LGm1 and LGm6 subtypes. Adding the double chromosome arm deletion status to the colour code of the PCA plots highlights a trend for the samples with these copy-number variations to behave as a whole more differently in relation to *IDH*-wild type than samples without these chromosome deletions. The strata formed by the two *IDH*-mutant groups of samples are, to a certain extent, superimposable with the strata formed by LGm2 and LGm3 samples. From these four plots, it can be seen that the DNA-methylation glioma classification adds extra levels of information into the molecular distinction of samples, in a way that is quite independent from tumour grade classification. This classification incorporates the *IDH* mutation status and 1p19q-codeletion information, and is still able to discriminate within *IDH*-wild type and -mutant samples the groups LGm1 and LGm6, which in fact show a very heterogeneous behaviour.

As for the similarities gene expression data have with the LGm classifier in terms of overall ability to discriminate glioma cases, as had been reported in the work of Ceccarelli and collaborators, in which a pan-glioma RNA-expression classifier was also developed (shown further down in Figures 3.5-3.6), it does not seem to capture the same levels of biological information, namely since it seems to be unable to distinguish any of the three *IDH*-wild type subgroups LGm3, LGm4 and LGm6. Alternative splicing separates the LGm subtypes very similarly to gene expression.

The third rows of panels in Figures 3.5-3.6 highlight two glioma transcriptomic classifiers: one based on microarray gene expression data of glioblastoma cases only (Verhaak et al., 2010), and the pan-glioma RNA-expression clusters classifier, developed in (Ceccarelli et al., 2016). Principal component 1 but mostly principal component 2 of gene expression separate quite clearly most Proneural (PN) subtype samples from Neural (NE) ones, independently of tumour grade. Alternative splicing principal component 2 again separates these two groups, although with more overlap, similarly to gene expression principal component 1. Also very interestingly, the four LGr pan-glioma RNA-expression clusters may be clearly observed in the gene expression PCA, along the first two principal

components, LGr 1 and 2 being two subgroups within the *IDH*-mutant-codel cases. Alternative splicing separates these groups as well as gene expression PC1 alone, and among its corresponding first six principal components neither reflected LGr1-LGr2 separations similar to the one obtained by gene expression PC2.

Finally, in the histology panels of Figures 3.5-3.6, histological types can be seen to intermingle, except for glioblastoma samples that separate well from the remaining samples, as had been observed in the grade panel. In the proteome cluster panels, both gene expression and alternative splicing can be seen to be able to separate samples from the two clusters identified by the reverse-phase protein lysate microarray platform.

In this work, there will be a focus on tumour grade, a strong prognostic marker associated with malignancy of the tumour tissue, and in particular on the pan-glioma DNA-methylation cluster classifier, as it is based on a quite stable epigenetic mark, is useful to classify many different clinically relevant subtypes, both in terms of prognosis and therapeutic management of patients. A better understanding of the LGm subtypes in terms of important cellular and molecular pathways of disease development and progression would be desirable. Alternative splicing regulation may play key roles in some of these pathways or others more generalized in gliomas.

In Figure 3.7, PCA scatter plots for individual alternative splicing event types are shown. Exon skipping plots are very similar to the plots for all events, due to the fact that this type of event is the most abundant: 10700 out of 17151 events. Compared to the other event types, skipped exon principal component 2 seems to form two more consistent LGm2 and LGm3 strata than alternative 3' splice site, alternative 5' splice site and retained intron event types. Mutually exclusive exons made this separation more similarly to skipped exons and were also able to separate transcriptome subtypes and pan-glioma RNA-expression clusters through principal components 1 and 3, in a way that was very similar to the combination of the two first principal components of gene expression (data not shown). Curiously, from the alternative splicing event types shown in Figure 3.7, it was the only one mapping samples according to grade along the principal component 1, like for gene expression, instead of along principal component 2. Although PCAs for alternative first exons and alternative last exons are not shown, these also separated samples from the two transcriptomic classifiers similarly to gene expression along their principal components 1 and 3 or 1 and 4, respectively. These data suggest that these three types of alternative splicing are more dependent on gene expression, an association that would not have been anticipated for mutually exclusive exons. As a final note, although alternative last exon events, similarly to what happened to gene expression and mutually exclusive exons, separated tumour grades along the principal component 1, the same did not happen with alternative first exon events. This might indicate that the ability to separate more clearly transcriptome subtypes is not related with the absence of the principal component of variance found for most alternative splicing event types, which is a dimension that separates samples in way that does not reflect any known clinical or molecular factor.

In order to try to assess if this principal component had a technical origin, a bias according to sample library size and sample source centre were checked for (Figure S2). In both cases, it was not possible to detect a source of bias, either through an accumulation of samples with a particular range of library sizes or with a particular source centre colour code on either side of PC1. Yet other possible factors behind alternative splicing first principal components were excluded, related with aneuploidy and mutational loads and quantification of tumour mass contaminations with cells from the immune system and stromal (general name for connective tissue) (Figure S3), which could be causing fundamental differences in the PSI quantifications. Information on these parameters was retrieved from (Ceccarelli et al., 2016) by separating the samples in two groups that approximately split the

variance along the principal component in two (principal component coordinates below and above 0.05) and looking at the eventual presence of mutually exclusive values taken up by each groups for the different variables. There were no mutually exclusive ranges of values between the groups, although there was a trend for a higher immune cell abundance and percent aneuploidy in the samples with a high PC1 score. It is possible that these parameters are related with the first principal component of PSI event types skipping exon, retained intron, alternative 3' splice site, alternative 5' splice site and alternative first exon, but this would have to be investigated further.

Finally, a very good separation between glioblastoma and low-grade glioma samples was obtained with alternative 3' splice site PSI values (Figure 3.7), which might be linked to particular regulatory requirements for all or a subset of these events and is worth a careful analysis.

3.1.3.1 Exploring the glioma alternative splicing signature

Results from the previous section suggested alternative splicing contrasted samples in parallel ways to gene expression. But the similarity with which these two levels of transcriptomic data had structured samples could be a consequence of the inclusion in the alternative splicing analysis of events that were highly dependent on their own gene expression.

In order to find if there was a glioma signature specific of alternative splicing, alternative splicing events whose PSIs were not significantly dependent on their own gene expression and whose range of PSIs varied to a reasonable extent across the samples cohort, so that differences between samples could be detected, were selected. These were then analysed by principal component analysis in order to identify main variance trends. Alternative splicing events with Spearman correlation FDR with their own gene expression above 0.01 and variance of 0.0225 (standard deviation = 0.15), as suggested for performing hierarchical clustering using PSI values in ("Percentage Splicing Index - Geuvadis MediaWiki," 2016), were selected, making up a total of 795 events.

Analysis of the first principal components associated with PSI values for these 795 events returned results that were very similar to the ones obtained using all non-zero variance 17097 events. Illustrating this point, two-dimensional scatter plots of sample mappings along PC2 and PC3 obtained using these two sets of alternative splicing events, with tumour grade and LGm colour schemes, are presented in Figures 3.8-3.9.

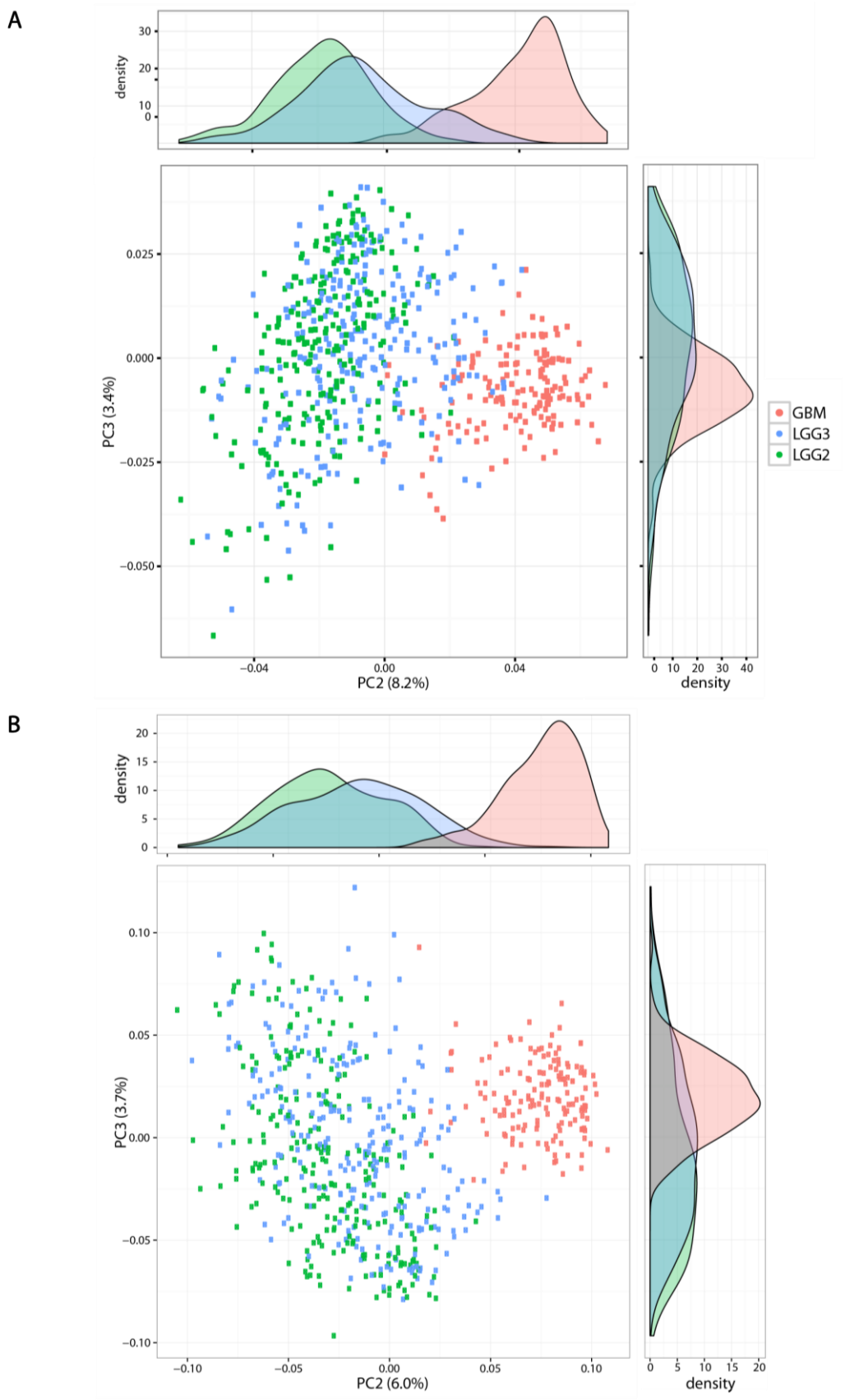


Figure 3.8 – Principal Component Analysis plots made on all measured AS events. (A) and on variable AS events whose regulation is not dependent of gene expression (B). Colours are according to tumour grade.

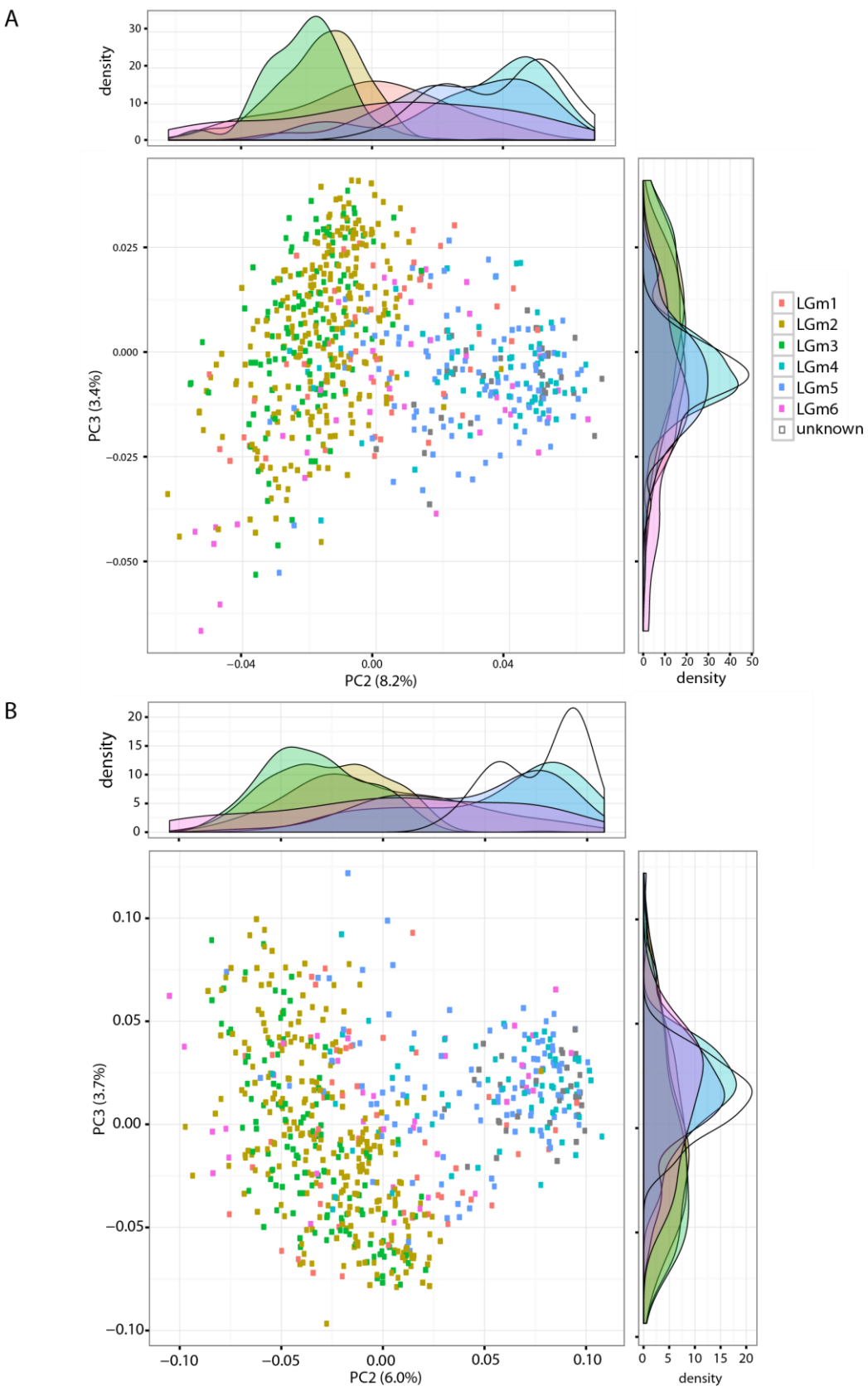


Figure 3.9 - Principal Component Analysis plots made on all measured AS events. (A) and on variable AS events whose regulation is independent of gene expression (B). Colours are according to DNA-methylation subtype.

3.1.4 Functional Analysis of the gene expression and alternative splicing malignancy axes

In the previous section it was shown that all gene expression and alternative splicing types had a principal component of variance along which samples from increasing tumour grade, and thus increasing malignancy, were mapped. A verification of whether these different axes ordered samples the same way was made, by looking at Spearman correlations between samples scores of the principal components of interest. Spearman's correlations obtained are shown in Figure 3.10.

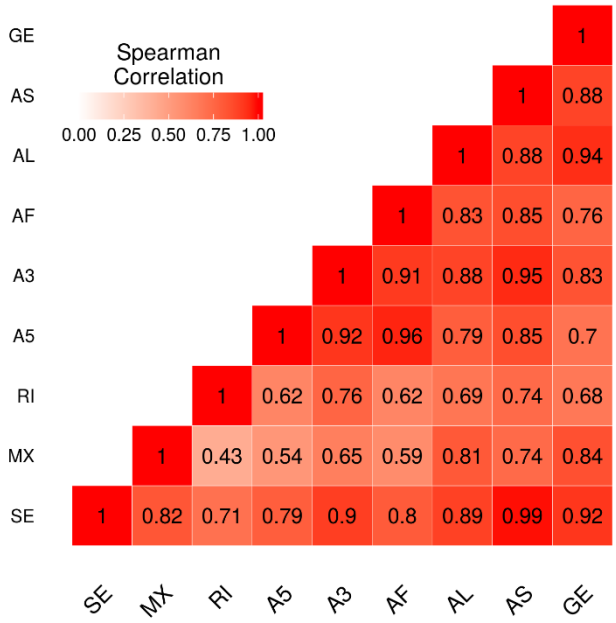


Figure 3.10 –Spearman's correlation coefficients for all pairwise comparisons of samples scores of malignancy-reflecting principal components using skipping exon (SE), mutually exclusive exon (MX), retained intron (RI), alternative 5' splice site (A5), alternative 3' splice site (A3), alternative first exon (AF), alternative last exon (AL), all alternative splicing events (AS) and gene expression (GE).

Sample rankings were indeed very similar, with only mutually exclusive exons and retained introns presenting higher dissimilarities. This observation suggests there are coherent changes in transcriptional outputs along this therefore single dimension of glioma malignancy.

In order to understand if there were coordinated gene expression and alternative splicing functionally relevant changes along these dimensions of variance, enrichment analysis on the gene expression PC1 and alternative splicing PC2 were performed for KEGG pathways, Reactome pathways and GO Biological Process terms. This was made by using the gene and event loadings from the principal components as the ranking metric for gene set enrichment analysis (GSEA) (Subramanian et al., 2005). This analysis would also help determine the extent to which gene classes defining the malignancy axes were coincidental in transcription and alternative splicing.

Amongst upregulated pathways in more malignant samples (positive enrichment scores), there were immune response, cell cycle, extracellular matrix organization or cell-signalling, all of them pathways known to be important for tumour cell proliferation and invasiveness (Figure 3.11).

Upregulated pathways in less malignant tumours included mostly neuronal cell lineage associated ones like neuroactive ligand receptor interaction, calcium signalling, long term potentiation, categories that likely reflect the fact that cells from lower grade tumours better resemble healthy glial cells, with neuronal functions preserved.

Other examples were proliferation associated cell signalling pathways like ERBB2, Wnt, MAPK and muscle contraction associated pathways, which correspond to gene sets that have many calcium channels, also known to play a major role in normal astrocyte function. It is possible that some of the alterations detected reflect rather a downregulation in high-grade tumours in relation to both lower grade ones and healthy tissues. These results were consistent with the ones from Wang and collaborators, who detected cell cycle and Wnt signalling as main up- and down-regulated pathways, respectively, in grade 4 *versus* lower grade astrocytomas (Z.-L. Z. Wang et al., 2015).

No gene sets were found enriched at an FDR cut-off of 0.05 for the splicing principal component 2 event loadings. This can probably be explained by the fact that, unlike gene expression regulation, which is known to happen frequently through coordinate changes of the different players of a functional cellular pathway (e.g. signalling, metabolic), alternative splicing regulation is known to operate less frequently this way. This difference frequently leads to the impossibility to get biological insights into the impact of alternative splicing changes using tools as gene set enrichment analysis. Still, GSEA performed on KEGG pathways gene sets returned the lowest FDR of 0.09 (p-value of 0.002) for the dilated cardiomyopathy gene set, containing genes involved in calcium channel transport (e.g. SLC8A1, ATPA2, CACNB4) and cell adhesion (e.g. ITGA6, ITGA3, LAMA2), biological functions that also appeared enriched in association with the gene expression malignancy axis (Figure 3.11).

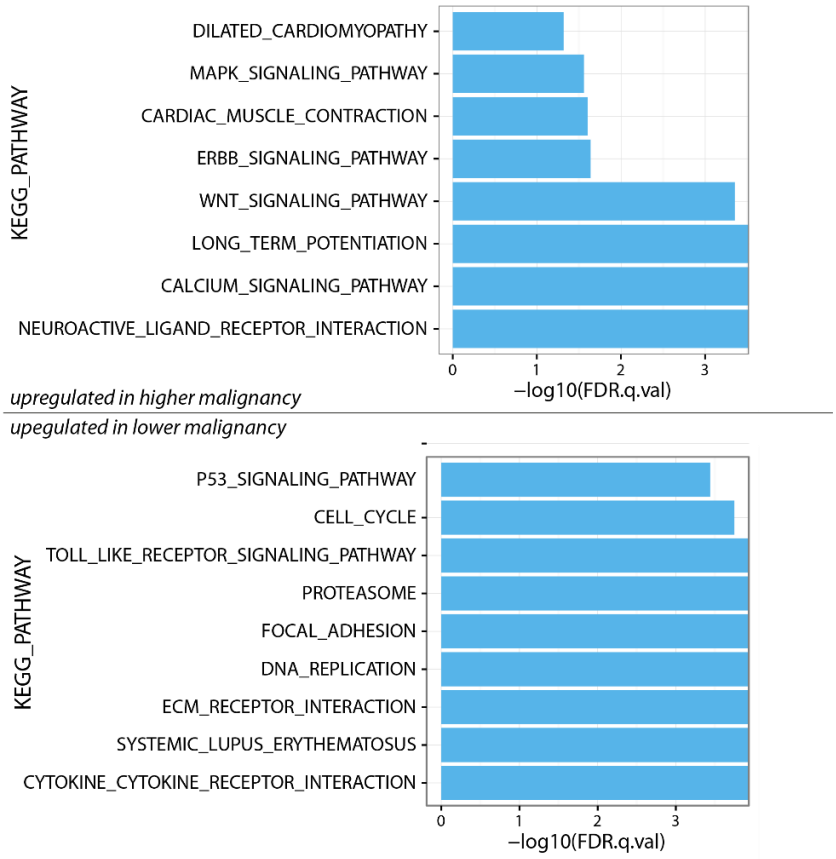
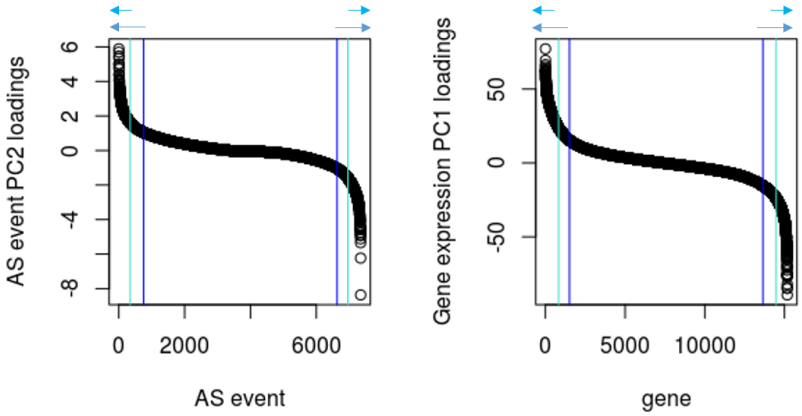


Figure 3.11 – Functional analysis of gene expression malignancy-reflecting principal component. Gene Set Enrichment Analysis using KEGG pathway gene sets on gene expression PC loadings was performed and selected functional categories at a FDR cut-off of 0.05 are shown.

Interestingly, enriched dilated cardiomyopathy genes were to a great extent different between gene expression and alternative splicing data sets. Namely, genes coding for different subunits of calcium transporters or integrins appeared either affected in terms of expression levels or of alternatively splicing, and members of the tropomyosin family (i.e. *TPM1, TPM2, TPM3*) appeared associated with different levels of malignancy only through alternative splicing changes.

Because most genes whose alternative splicing varied the most along PC2 did not fit into functional categories represented by a given gene set, this analysis did not help in understanding if there was a great level of overlap between these and the genes contributing more to the gene expression PC1. In order to assess this question, the enrichment of alternative splicing events with high absolute value loadings among expressed genes that also had loadings with high absolute values was tested (Figure 3.12).

Although accounting for the 20 % of genes with more variance in alternative splicing and expression resulted in a significant enrichment at an α of 0.05, when only the 10 % genes that contributed more for variance across the two principal components were considered, this enrichment ceased being significant. These results attest for the conclusion that both gene expression and alternative splicing variation across glioma cases of different degrees of malignancy reflect important cell identities and functions, which are partially redundant between the two, but also partially exclusive.



Hypergeometric test:

Quantile 0.80
p-val = 0.02

	HighAS loading	LowAS loading
HighGE loading	169	568
LowGE loading	1301	5310

Quantile 0.90
p-val= 0.23

	HighAS loading	LowAS loading
HighGE loading	26	197
LowGE loading	709	6416

Figure 3.12 – Alternative splicing events and transcribed genes with higher loadings across the malignancy axis affect different sets of genes. Hypergeometric tests to evaluate the enrichment of genes affected by alternative splicing events that contribute the most to PC2 among the transcribed genes that also contribute the most to the gene expression PC1 were performed. Loadings thresholds corresponding to quantiles 0.80 and 0.90 were used for selection of alternative splicing events and transcribed genes with high contribution to the two principal components to enter in the statistical test.

3.1.5 Analysis of differential gene expression across DNA-methylation cluster subtypes

Before moving on into the identification of alternative splicing events regulated differently in the six glioma DNA-methylation groups, differential gene expression analysis between the same groups was performed. Knowing the genes more significantly altered across those conditions would be

interesting for comparison with the list of more differentially spliced genes. Furthermore, this analysis could allow to detect splicing regulators relevant in particular glioma subtypes.

Differential expression analysis was carried out using the *edgeR* Bioconductor package, through fitting negative binomial generalized linear models for each gene and performing ANOVA F-tests for differences between the LGm groups. From a group of 15957 annotated genes consistently detected in glioma samples, 5970 appeared differentially expressed, at an F-test FDR cut-off of 0.01 and a minimum log2-fold change in expression of 1 between any of the LGm groups 1,2,4,5 and 6, and the least malignant glioma group LGm3.

Among these genes, there were 183 established cancer driver genes, identified by comparison with the database from The Cancer Gene Census project (Forbes et al., 2015; Futreal et al., 2004), which is a curated database that stores information about genes frequently mutated in cancer shown experimentally to play active roles in disease progression. Among these are *ERBB2, EGFR*, *CDKN2C*, *MDM2* and *TET1* oncogenes, whose implication in glioma development has been thoroughly shown (Brennan et al., 2013; Suzuki et al., 2015).

Apart from the core spliceosome components that are required to carry out the splice site recognition and subsequent catalysis, there are trans-acting splicing factors that assist in the selection of particular splice sites according to the cell type or extracellular signals received at a particular time. Comparing the list of differentially expressed genes between DNA-methylation subtypes with a list of RNA-binding proteins (RBPs) and splicing factors obtained from (Sebestyén, Singh, et al., 2015), 195 RBPs and 41 splicing factors (Figure 3.13) were found. At the top of the list were IGF2BP2 and IGF2BP3, insulin-like growth factor 2 mRNA-binding proteins 2 and 3, which are cancer and type II diabetes risk factors (Schaeffer et al., 2010; Zhang, Chan, Peng, & Tan, 1999). Their better characterized function involves binding to the insulin-like growth factor 2 and the cell adhesion protein CD44 UTRs in order to regulate transcript stability and translation (Nielsen et al., 1999) but they have also been implicated in splicing (Cleynen et al., 2007). These two splicing factors are lowly expressed exclusively in LGm2 and LGm3, the two *IDH*-mutant subtypes having higher levels of DNA-methylation.

At the criteria used for differential expression classification, the only splicing factor already known to be implicated in glioma was *A2BP1* (*RBFOX1*), whose downregulation in glioblastoma has been shown to compromise the differentiated cell state, known to be lost in cancer cells (Hu et al., 2013). Curiously, among the *IDH*-wild type samples, this transcription factor appears downregulated in LGm4 and LGm5 but upregulated in LGm6 samples, where almost 8 times more mRNA is expressed in relation to the two other groups.

PTBP1 and PTBP2, which promote proliferation and cell migration in glioma cell lines (Cheung et al., 2009), were nevertheless also consistently differentially expressed between LGm groups (FDR < 4.3×10^{-40} and 2.7×10^{-28}), although having fold changes lower than 2 in relation to the LGm3 subtype. Interestingly, while PTBP1 was upregulated in LGm4, LGm5 and LGm1, PTBP2 was downregulated in LGm4 and LGm5, while keeping the upregulation in LGm1, always in relation to LGm3.

From observation of Figure 3.13, other patterns of expression of splicing factors between LGm groups can be found, with PCBP3 exhibiting a descending expression gradient according to the order LGm3, 2, 1, 6, 5, 4. A similar trend happened with PABPC5. Another interesting pattern of expression across groups is the one of KHDRBS2, which shows very low levels of expression for LGm4 and 5 subtypes and higher for the two *IDH*-mutant subtypes without 1p-19q codeletion LGm1 and 2.

Importantly, 13 of the differentially expressed splicing factors have known RNA-binding motifs and are thus good candidates for alternative splicing regulation through direct interaction with splicing enhancer and silencer sequences.

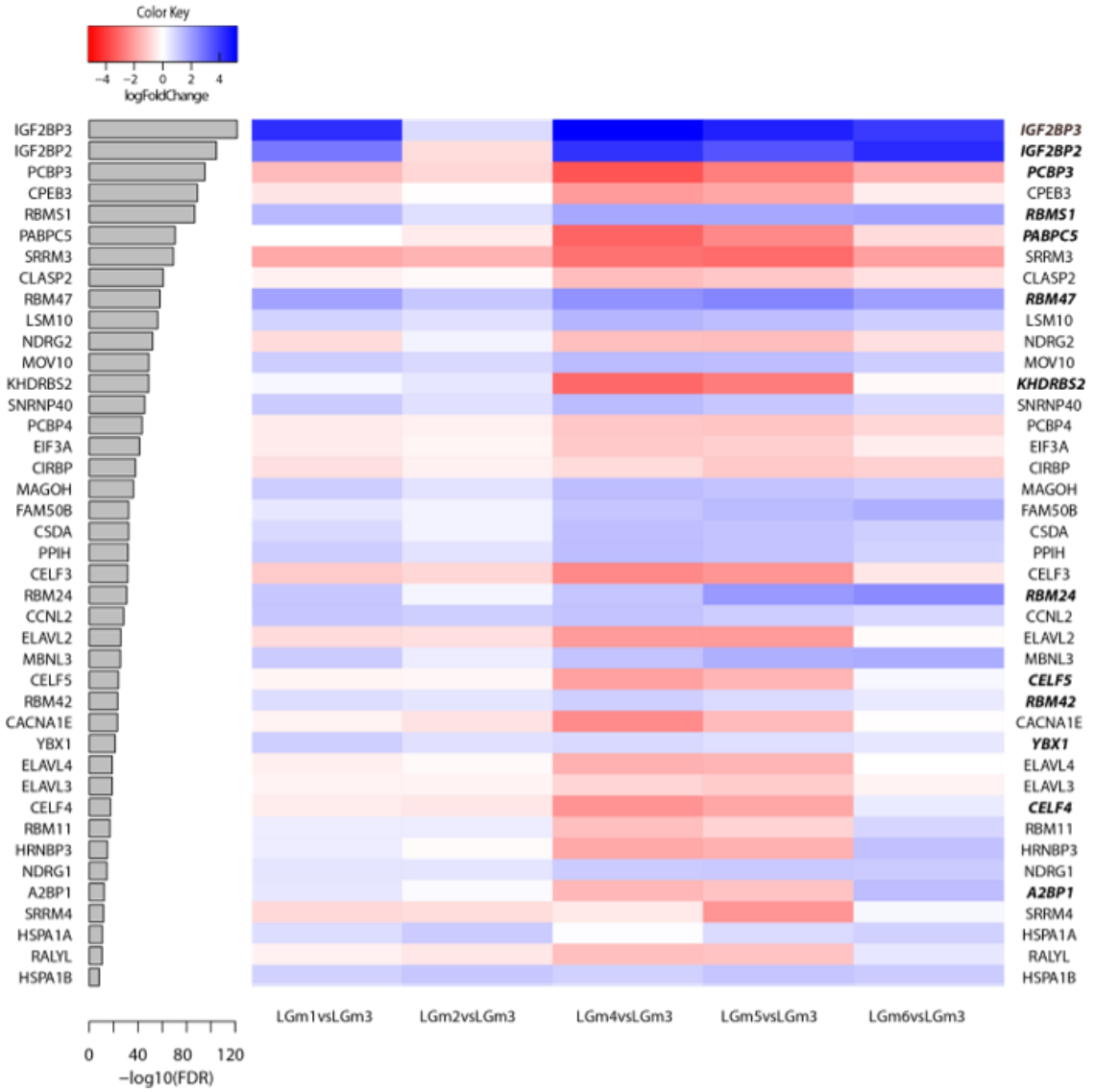


Figure 3.13 - Differential expression statistics and relative expression levels of known splicing factor genes across glioma DNA-methylation subtypes. Genes that code for proteins with known RNA-binding motifs are shown in bold.

3.1.6 Analysis of differential splicing across DNA-methylation cluster subtypes

In order to determine which AS events appear differentially spliced between the DNA-methylation cluster subtypes, a Kruskal-Wallis test was run on the PSIs of the 17097 events having variance different from zero, across the six LGm groups of samples. This analysis rendered a total of 10507 AS events showing differential splicing in at least one of the DNA-methylation cluster subtypes at an FDR < 0.01. This is a quite high number of significant hits, and may result from the limitations of a non-parametric test to deal with highly heterogeneous PSI variances shown by the different LGm groups. In order to check if the use of a significance value of 0.01 is indeed a good choice, six events were chosen from the top and bottom of the Kruskal-Wallis FDR-ranked list and their PSI distributions plotted (Figure 3.14).

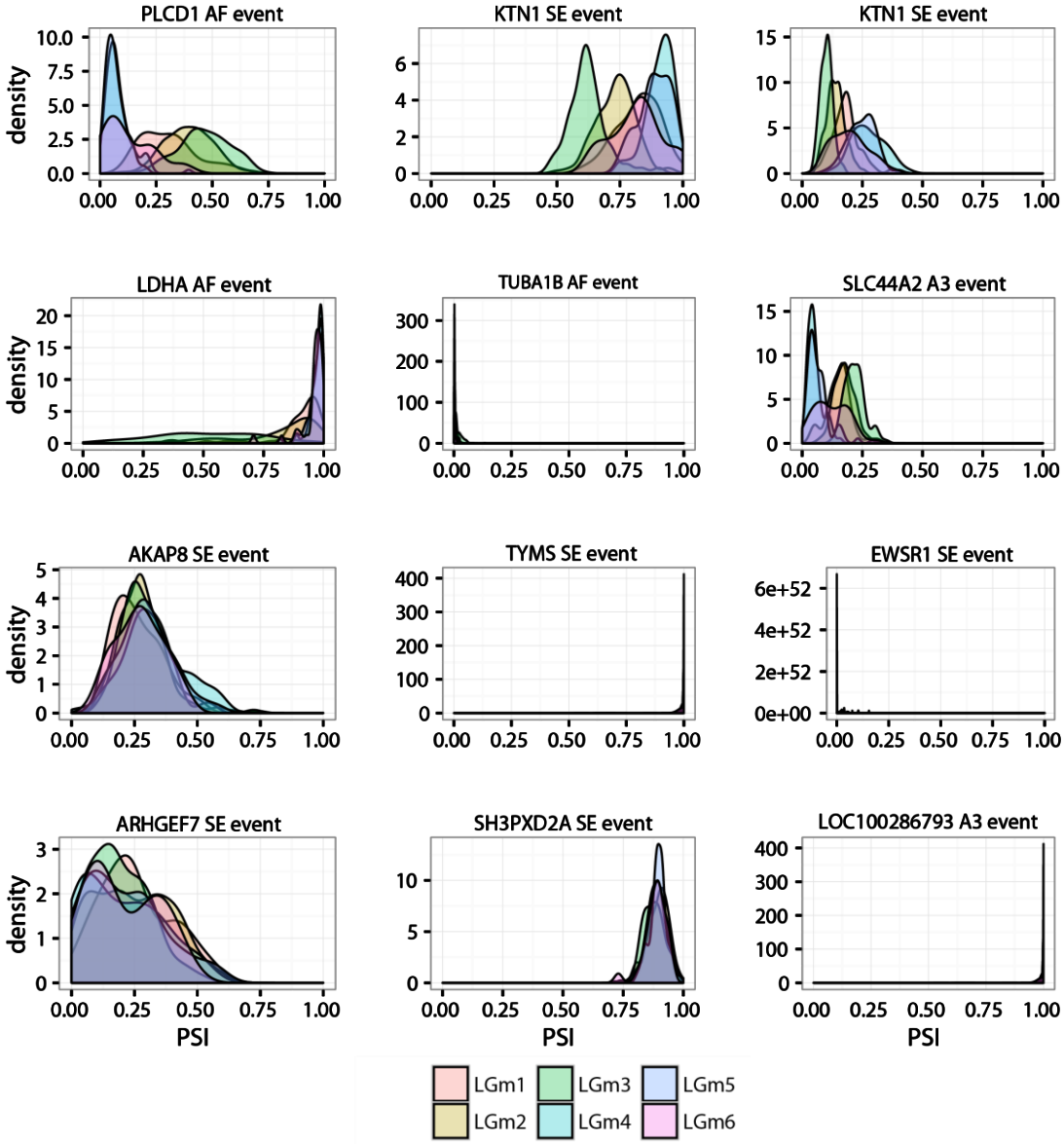


Figure 3.14 – PSI distributions for 12 alternative splicing events that appear differentially expressed across DNA-methylation subtypes, at a Kruskal-Wallis FDR significance of 0.01. The upper six plots are the top significant events (FDR < 2.0 x 10^{-74}), while the bottom six are the least significant (FDR close to 0.01).

While the first six alternative splicing events show clearly distinct PSI distributions between certain LGm groups, the bottom six splicing events show either having largely superimposed PSI distributions from all groups or very narrow PSI distributions with some outliers.

In order to make a better selection of differentially spliced events across LGm groups, the variance threshold of 0.0225 used before in the exploratory analysis was employed again. Using this event filter, it was possible to identify AS events able to distinguish DNA-methylation cluster samples at a Kruskal-Wallis FDR below 1 x 10^{-9} (Figure 3.15), even though large superposition of individual LGm PSI distributions was still observed.

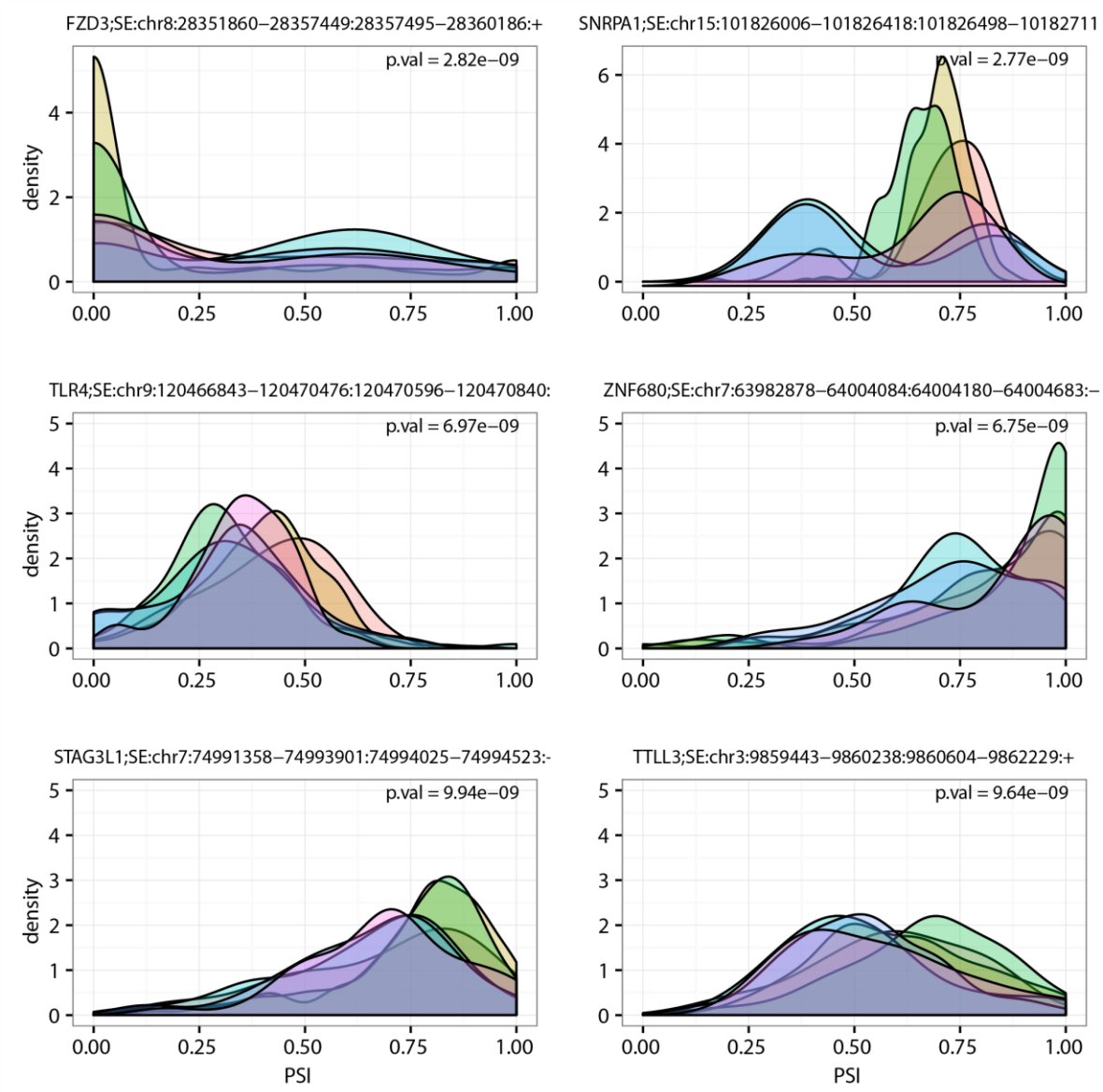


Figure 3.15 – PSI distributions of six AS events that just the criteria to be considered differentially spliced between glioma DNA-methylation clusters. Events were chosen from the 20 having the highest FDR in the group of 721 having PSI variance higher than 0.0225 and Kruskal-Wallis FDR below 1 x 10^{-9}. Colours are as defined in Figure 3.14.

To evaluate the criteria used to consider an event to be differentially spliced, a nearest shrunken centroid method, PAM (Tibshirani et al., 2002), was used to create a DNA-methylation classifier from the glioma PSI data. The idea behind using this supervised classification algorithm was to compare the variance and Kruskal-Wallis FDR ranges of the alternative splicing events selected for LGm class distinction with those previously chosen as thresholds for differential splicing across LGm subtypes.

A first classifier built to distinguish all six DNA methylation subtypes was created, getting an overall cross-validation error rate of 0.34 and yielding 2347 classifying AS events. Consistently with the high error rate and what had been noticed earlier by PCA, its cross-validation class prediction plot showed the impossibility to accurately identify samples from DNA-methylation clusters LGm1 and LGm6 based on PSIs (Figure 3.16A). A classifier excluding samples from these subtypes performed better in cross-validation, exhibiting a test error of 0.26, using 1397 AS events (Figure 3.16B).

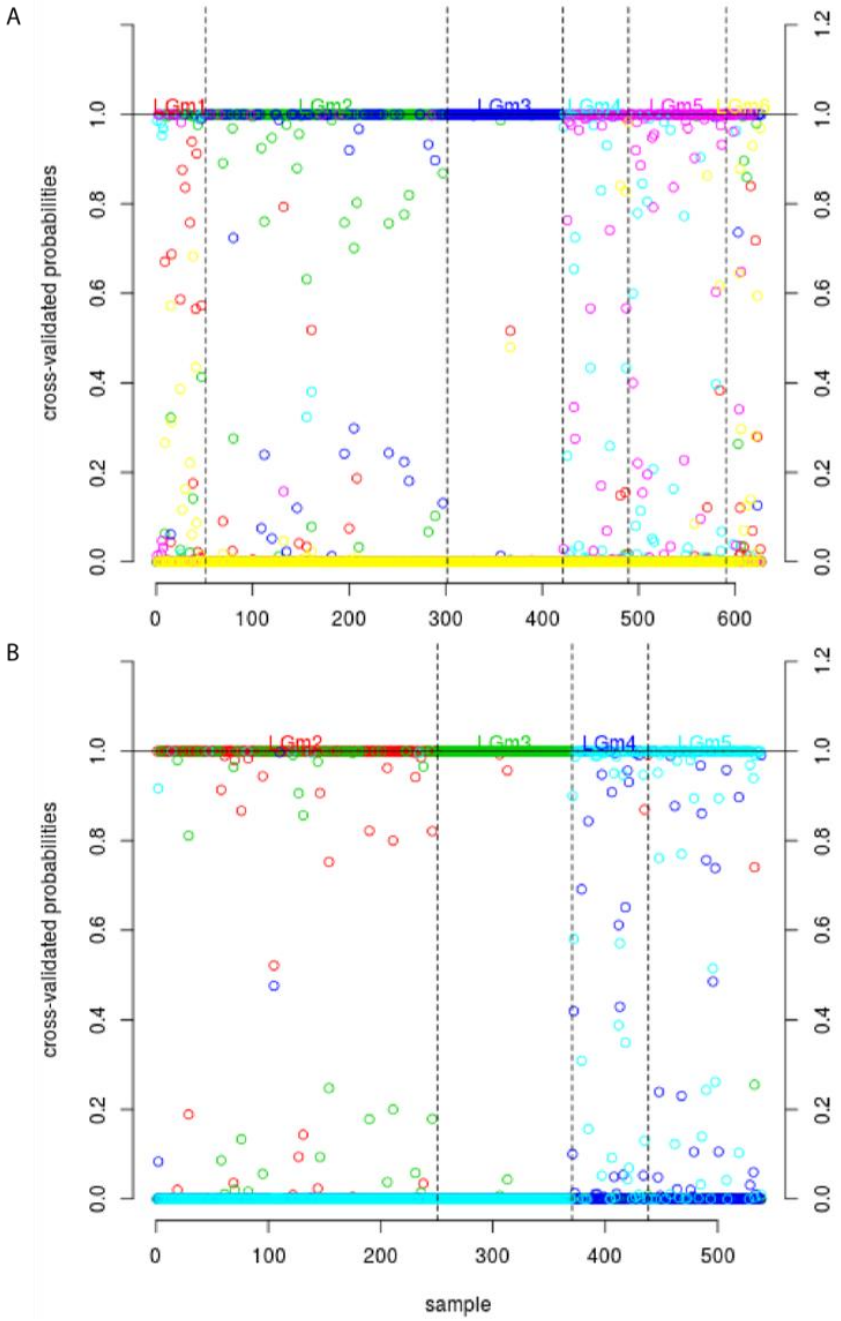


Figure 3.16 – Plots for cross-validation of two supervised classifiers produced with PAM algorithm, one for the six LGm1-6 subtypes (A) and another for subtypes LGm2-5 (B). Cross-validation probabilities refer to class predictions applied each of the training sets. Both classifiers perform better for LGm2, LGm3 and Lgm5 subtypes.

Although 90% of the PAM classifier events meet the empirically chosen Kruskal-Wallis FDR cut-off of $1x10^{-9}$, much less consensus between variance ranges was found. Indeed, the use of a minimum variance threshold of 0.0225 would have kept more than 70 % of the PAM classifier alternative splicing events from being considered differentially spliced across DNA-methylation subtypes (Figure 3.17).

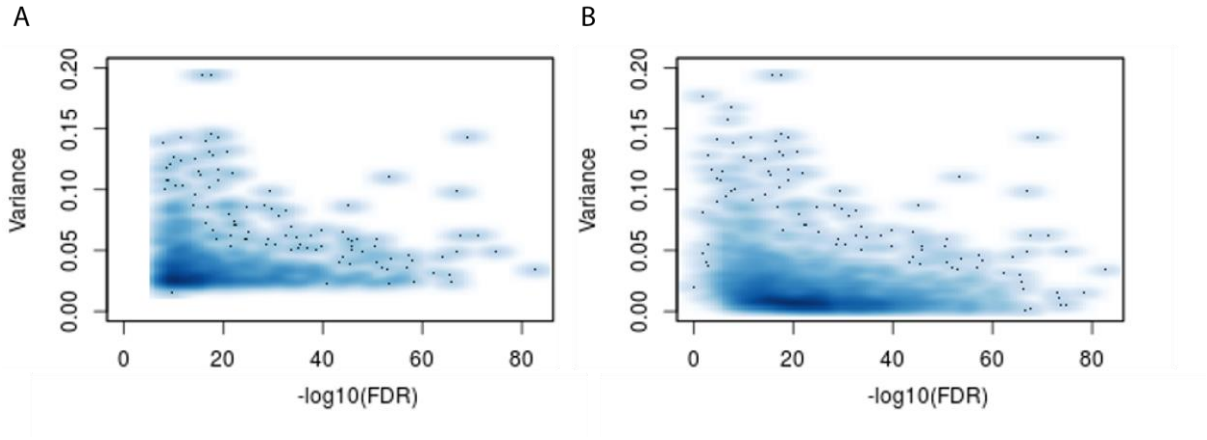


Figure 3.17 - Variance and Kruskal Wallis FDR of alternative splicing events that vary across DNA-methylation clusters. (A) Scatter plot relative to alternative splicing events considered differentially spliced according to variance and Kruskal Wallis FDR thresholds found adequate through visual exploration of events PSI distributions (FDR<1x10^{-9} and variance >0.0225). (B) Scatter plot relative to alternative splicing events considered to accurately distinguish DNA methylation clusters via PAM closest shrunken centroid classifier.

These observations allowed to conclude that the criterion of choosing a minimum variance threshold for differential splicing classification across DNA-methylation groups had limitations. An alternative criterion for the definition of differential splicing was used instead, which consisted of setting a minimum value for the difference of median PSI values between at least two LGm groups. The minimum value was 0.1, commonly used in published studies profiling differential splicing between two conditions.

A total of 1762 alternative splicing events met the criteria of having a Kruskal-Wallis FDR below 1 x 10^{-9} and a minimum 0.1 median difference between two LGm groups. Remarkably, from these events, 1395 happened in a group of 1058 genes that are not differentially expressed among the same six glioma subtypes. This observation, together with the finding that most alternative splicing events lowly correlate with the expression of their cognate gene, provide a clear indication that there is a large group of alternative splicing events that are being regulated independently from rates of RNA polymerase II transcription, the main known mechanism dictating dependence of splicing rates on transcriptional output.

Similar to what had been done with gene expression data, an identification of the number of differentially spliced events among glioma methylation subtypes affecting known oncogenes or tumour suppressor genes was made, using information from The cancer Gene Census project database. This comparison allowed to detect 105 differentially spliced events (corresponding to 73 genes), from which 89 corresponded to transcripts from 64 genes that are not differentially expressed across DNA-methylation subtypes (*see* Table 3.1 for a detailed summary of this information).

Table 3.1 – Number and role in cancer of genes and AS events differentially expressed across glioma DNA-methylation subtypes.

	Role in Cancer				
Feature Type	unknown causality in cancer	implication in cancer unkown	oncogene	oncogene/TSG*	TSG*
DGE	5787	126	35	8	14
DAS	1657	78	6	2	19
DASnotDGE	969	50	4	1	9

* TSG – tumor suppressor gene

Particularly, in what concerns alterations in splicing factors, whereas among the list of genes considered to be differentially expressed these were 41, among the differentially spliced genes there were 46 splicing factors were found to be differentially spliced but none of these was among the 41 differentially expressed ones. Four of the differentially spliced splicing factors are also known to carry cancer driver mutations: HNRNPA2B1, U2AF1, RBM10, FUS. The first two were significant through the-Kruskal-Wallis test and their PSI distributions are shown in Figure 3.18. The alternative splicing event involving HNRNPA2B1 gene was an exon whose skipping generates a protein isoform known as hnRNPA2, lacking residues 3 to 14 at the its beginning, which encode a nuclear localization signal (information source: Uniprot database). Interestingly, this protein has apart from its splicing function been shown to shuttle actively between the nucleus and the cytoplasm to transport mRNAs to particular cell sites, being particularly important in oligodendrocytes (Munro et al., 1999). Its prevalent localization in the cytoplasm has also been shown to be an early lung cancer biomarker (Nichols et al., 2000). The inclusion of the also differentially spliced exon 3 of U2AF1 causes a change of seven aminoacids in the protein's N-terminal portion that allows heterodimerization with U2AF2 to assist the function of 3' splice site selection. This isoform, also termed U2AF35b, was reported to make this interaction less efficiently (Pacheco et al., 2004), with potential consequences for splicing regulation. Association of U2AF1 with cancer comes from studies where mutated forms of this protein in acute myeloid leukaemia were shown to cause abnormal splicing in cell-cycle regulator genes and RNA processing genes mutated in different cancers (Przychodzen et al., 2013).

Another two of the most differentially spliced events across LGm groups affecting splicing factor genes involved *PCBP2* and *HNRNPD.* Both events (Figure 3.18) involve the generation of alternative protein isoforms of these Poly(rC)-binding protein and hnRNP family protein whose specific functions are not known. PCBP2 is involved in the control of innate immune response (You et al., 2009) and hnRNPD (Yoon et al., 2014) in increasing or decreasing the steady state of mRNA molecules, with important implications in genome integrity maintenance.

Many of the alternative splicing events found in the literature to be relevant in glioma were found to be differentially spliced in this study, confirming their general interest in the context of the disease, namely in differentiating between glioma subtypes. One such example is *RTN4* exon 3 inclusion, known to be regulated by the splicing factor PTBP1 and found to be preferentially included in LGm2 and LGm3 and the least in LGm4 and LGm5. Exon 6 of *ANXA7*, known to be preferentially included in the brain, had a median PSI below 0.1 for LGm1,4,5,6. PSI medians for the mutual exclusive event involving *TPM1* exons 5 and 6 got close to zero in LGm4,5, and higher than 0.5 in other subtypes. *FGFR1* exon 3 was preferentially excluded in LGm4,5, with the resulting predicted increase of FGFR1 signalling. *EGFR*, in turn, exhibited the highest expression in LGm4 and the least in LGm6, but no differential splicing was found related to it. Finally, exon 8 of the tumour suppressor gene RECK was specifically excluded in LGm4.

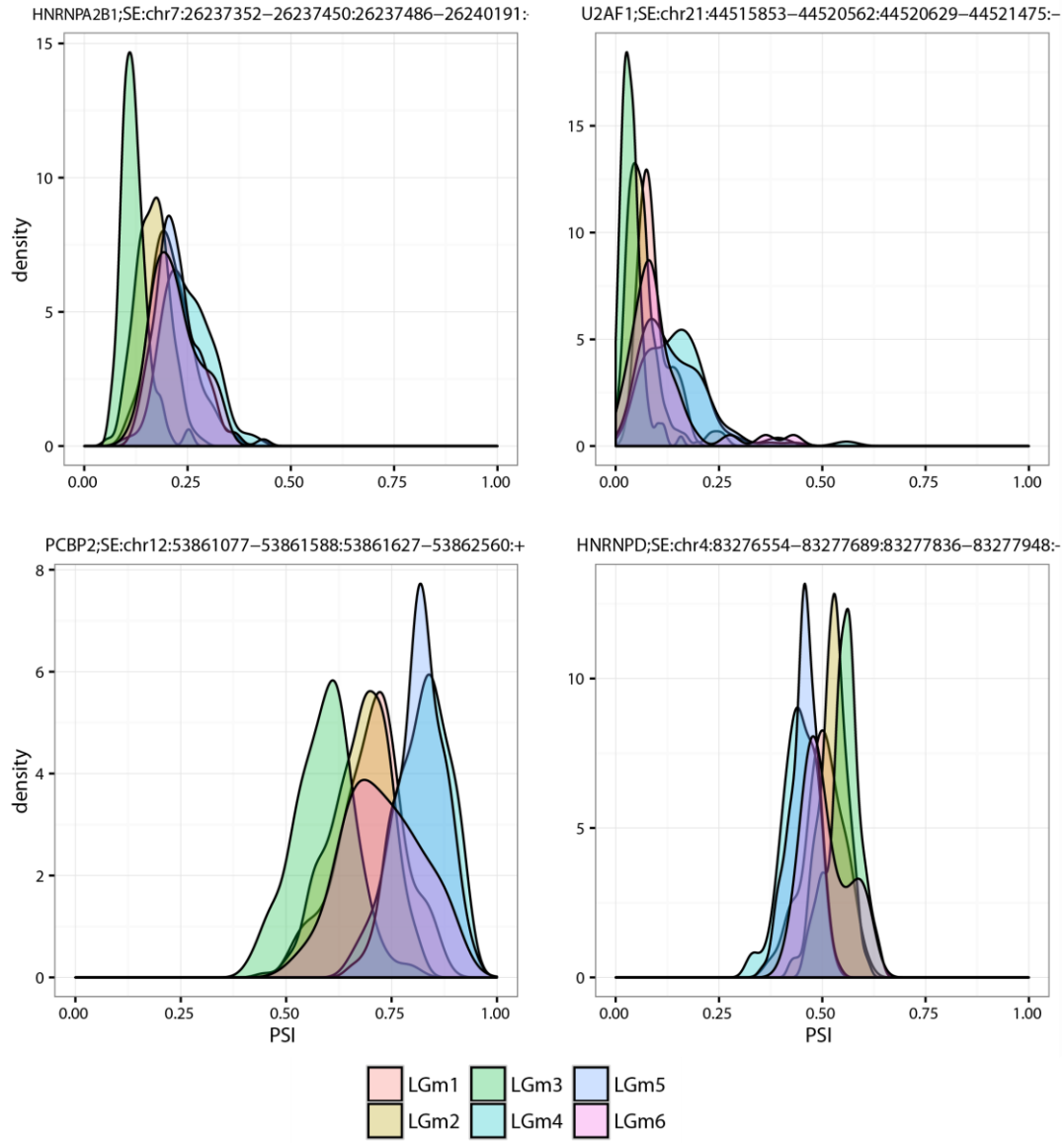


Figure 3.18 – PSI distributions of four alternative splicing events that affect splicing factor genes. Colours are according to DNA-methylation subtypes.

3.1.7 Functional Analysis of gene expression and alternative splicing changes in LGm subtypes

In order to identify biological processes enriched among differentially regulated alternative splicing events, GSEA was done for Reactome and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, as well as Biological Process (BP) gene ontology (GO) terms-analysis. This analysis returned very interesting information to assist in the study of splicing changes in glioma. In parallel, a similar analysis was run for the differential gene expression data. Gene sets found enriched at an FDR adjusted p-value cut-off of 0.05 were considered. Significantly, no overlap was found between the sets enriched in differentially expressed and differentially spliced genes (Figure 3.19-3.20).

Cell pathways and biological processes enriched among differentially expressed genes involved cell adhesion, immune response, p53-signalling pathway and pathways affecting certain cancer types.

As for alternative splicing, there were fewer gene sets returning statistically significance. While no cancer-specific pathways appeared affected, some neuronal pathways did.

KEGG pathways, GO BP terms and Reactome pathways associated with RNA processing and specifically the spliceosome were enriched in differentially expressed genes, consistently with what had already been reported for colon adenocarcinoma and breast cancer (Danan-Gotthold et al., 2015). Genes having their transcripts ratios affected include many encoding ribosomal proteins (e.g. *RPL10, RPS15, RPL38, RPL13A, RPSA, RPS15A*), genes related to nonsense mediated mRNA decay (e.g. *SMG7*) or *YWHAZ* that encodes the 14-3-3 protein zeta/delta, a protein implicated in many signal transduction pathways and having known roles in mediating apoptotic pathways (Nishimura et al., 2013).

Among the enriched Reactome pathways, there were two related with protein metabolism regulation (involving genes like *TUBA1B, RPL7, TUBA1C, EEF1B2, EEF1G*), signal recognition particle (SRP) dependent targeting of membrane and secretory proteins to the cell membrane, involving again ribosomal genes like *SEC11A, RPN2, SSR2*. Interestingly, genes coding for apoptotic proteins was one, (like *CTNNB1, TJP2, MAPT, ACIN1, DFFA*) are also enriched among those having splicing changes across DNA-methylation glioma subtypes. KEGG mitochondrial pathways were found enriched among genes with altered splicing across gliomas. So were Parkinson's disease gene sets that include *NDUFV3, PARK7, ATP5J*, which code for nicotinamide adenine dinucleotide (NAD) and adenosine triphosphate (ATP) conjugated proteins. Most of the genes in this set ranking high for differential splicing were also differentially expressed though. As for alterations in Alzheimer's disease associated genes, those contributing the most to the enrichment score are shared with the Parkinson's enriched gene set. So are *BID*, *APP* (the amyloid beta precursor protein, which appears to be specifically affected at the level of its alternative splicing and not in terms of overall gene expression), protein phosphatases PPP3CA/C genes or the important proliferation/cell-differentiation signal-transduction mediator MAP3K. As for the pathogenic *Eschericia coli* infection pathway hit, the genes that contributed to the enrichment score had many different functions. They could be 1) cytoskeleton associated proteins genes, like *TUBA1B/C, TUBB6/3* that are brain-specific tubulin variants, and actin-related *ACTB, ACTG1, ARPC4* genes, 2) genes associated with cell-adhesion functions (e.g. *CTNNB1, ITGB1*), 3) signalling transduction (e.g. YWHAZ, FYN1, ARHGEF2) and finally, 4) a very important cell-cycle regulator encoding gene: CDC42. Still about this gene set, the gene classes that appeared to be exclusively affected at the level of alternative splicing were the actin related and signalling ones.

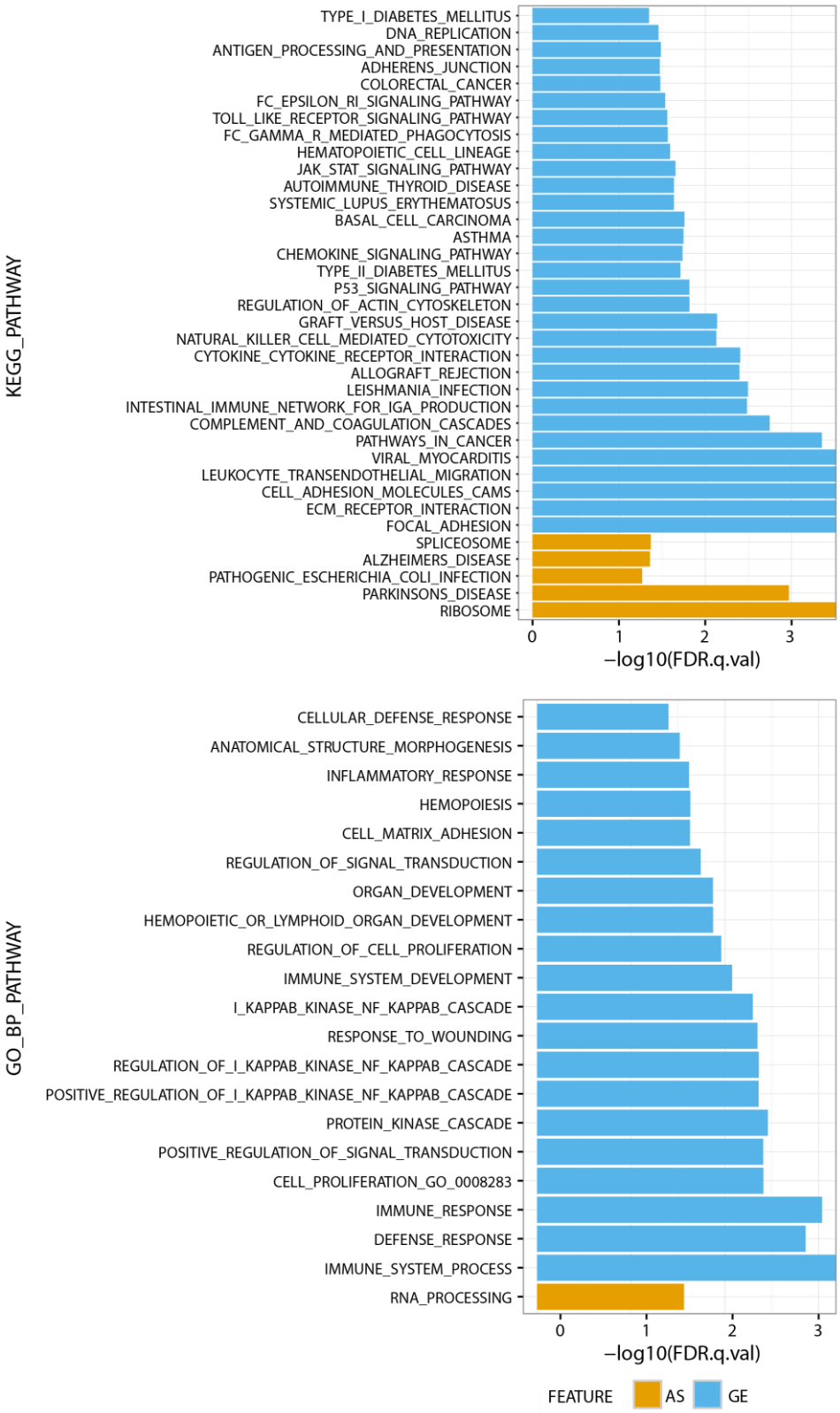


Figure 3.19 – Biological pathways and cellular processes enriched among differentially spliced and differentially expressed genes. Significant KEGG pathways and GO Biological Process terms obtained by GSEA for alternative splicing (AS) and gene expression (GE) are shown.

Figure 3.20 – Biological pathways and cellular processes enriched among differentially spliced and differentially expressed genes. Significant Reactome pathways obtained by GSEA for alternative splicing (AS) and gene expression (GE) are shown.

3.2 INVESTIGATION OF THE VALUE OF ALTERNATIVE SPLICING IN GLIOMA PROGNOSIS

Various factors act together in determining disease progression. In the case of glioma, tumour grade, age of the patient and LGm subtype, a molecular classifier that incorporates the information of other molecular risk factors that have been known for some time: *IDH* mutation status and *1p19q* chromosome arms deletions, are the most important factors to predict patient outcome.

The LGm classification system is quite recent and much is still to be discovered in terms of the specificities of these subtypes, namely in terms of regulation of alternative splicing. Discovery of LGm subtype markers that impact more on the disease, or that can at least be useful for diagnosis, for example in the absence of DNA-methylation data, can be quite important.

Furthermore, alternative splicing regulation could still be informative in terms of prognosis beyond the known risk factors. Using the follow-up information from the patients of the studied cohort, in particular the overall survival, the prognostic value of alternative splicing was studied.

3.2.1 Prognostic value of gene expression and alternative splicing malignancy axes

Gene expression and alternative splicing organized patients very similarly along the main principal component of explained variance that grouped the samples consentaneously with clinical indicators and established/published molecular signatures. In particular, these principal components organized the samples along a gradient of malignancy, which separated better grade 4 from grade 2 samples, as opposed to grade 3 samples that localized mostly in between.

Glioma histological grade dictates very different prognosis, as can be observed in Figure 3.21.

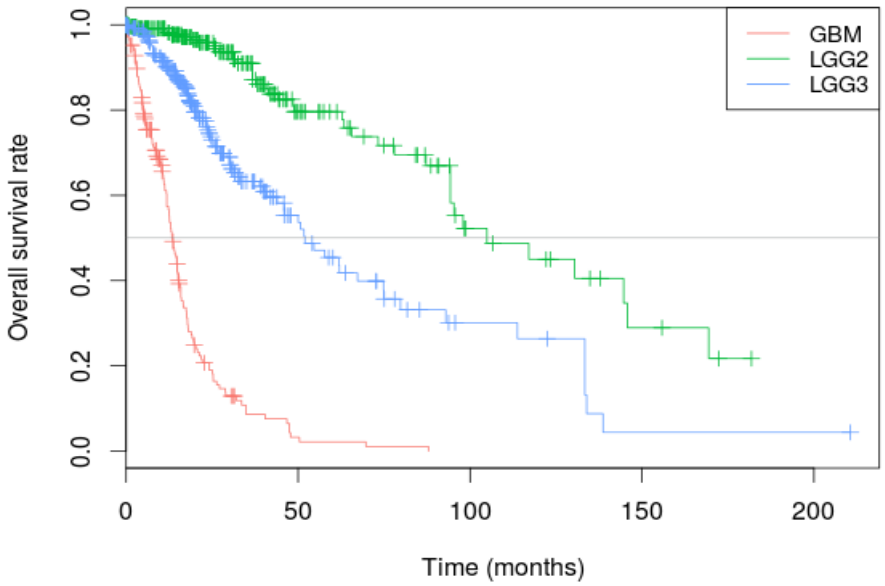


Figure 3.21 – Survival curves for different WHO grade gliomas. . Kaplan-Meier curves for grade IV GBM patients, and grades II (LGG2) and III (LGG3) patients, show three clearly distinct prognosis groups.

It then became interesting to test if the principal component 2 of the whole alternative splicing data, as well as the gene expression principal component 1, could indeed reflect a gradient of malignancy, even better than WHO grade classification in predicting patient outcome. Cox proportional-hazards models were run both on the sample scores of the referred principal components and on the samples' WHO grade strata, using overall survival as the event of interest. The general equation for performing the regression was:

$$h(t) = h_0(t)\exp(\beta_{x1}\, x\, \mathrm{x1}),$$

where $h_0(t)$ is the baseline hazard at time t, β_{x1} and x1 are respectively the regression coefficient and values taken by the independent explanatory variable x1 ("Fundamentals of Biostatistics 7th edition (9780538733496) - Textbooks.com," 2016)(*see* Methods).

Then, the amount of information about patient's survival taken from the model was assessed using the concordance index, an indicator that is part of the output from the *coxph.fit* function from the Bioconductor package *prodlim*, and consists on a goodness-of-fit test that compares the ranks of the survival time of the patients with the ones predicted by the explanatory variable being evaluated (Harrell, Califf, Pryor, Lee, & Rosati, 1982). The results are presented in Table 3.2. Cox proportional-hazards model for alternative splicing data PC2 got a higher concordance index than the Cox model based exclusively on WHO grade, as much as gene expression PC1 did, which showed the value of transcriptional programs associated with these components of variance to predict patient outcome in a finer way than tumour grade does.

Table 3.2 – Cox proportional-hazards models for malignancy-reflecting variables. Hazard Ratio (HR) with 95 % confidence interval (95 % CI) for patient overall survival according to different factors. Number of events: 239 out of 659 patients. p-values shown are for the log-rank test and relative to the HR estimate of each variable, being all significantly against the null hypothesis of the variable not having an effect on survival (α = 0.001). HR – Hazards Ratio; CI – Confidence interval.

Variable	Levels	HR	95 % CI	p-value	Concordance index
Alternative Splicing PC2	Sample score	1.19×10^{17}	9.15×10^{14}-1.54×10^{19}	< 0.001	0.80
Gene Expression Splicing PC1	Sample score	0.133	0.104-0.17	< 0.001	0.81
WHO Grade	III	3.21	2.15-4.78	< 0.001	0.78
	IV	19.8	12.9-30.3	< 0.001	

3.2.2 Prognostic value of individual genes and AS events

With the aim to discover alternative splicing events and genes that could work as good glioma prognostic markers, Cox proportional-hazards models were run for each alternative splicing event and gene. In addition, since there was the interest to distinguish the relative strength of splicing events in predicting patient's overall survival as compared to the expression of their cognate gene alone, so as to get insight into the clinical relevance of particular regulated exons, additive Cox regression models taking into account gene and alternative splicing event as explanatory variables were also run (*see* Methods).

At a level of significance of 0.01 FDR, a very high number of genes and alternative splicing events were found to be informative about patient's overall survival: 11794 out of the 15189 genes, 6657 out of 17097 alternative splicing events (2011 affecting 1204 genes that are not themselves prognostic markers) and 5991 alternative splicing events in the group of models adjusted for gene expression levels (1767 affecting 1072 genes that are not themselves prognostic markers).

Distributions of concordance indexes for the statistically significant markers from this survival analysis are shown in Figure 3.22, in which it becomes clear that there are potentially good individual markers among both alternative splicing events and expressed genes, showing concordance indexes above 0.6. Although this value is lower than the ones found previously while testing for the prognostic value of the principal components, it still tells about the potential of these individual transcript as malignancy markers. Effectively, the presence of high numbers of "expressionally prognostic genes" in glioma in comparison to other types of tumour had been previously reported in a pan-glioma study on cancer prognostic genes from RNA-seq data (Anaya, Reon, Chen, Bekiranov, & Dutta, 2016). In this study, it is pointed out that parameters like cohort size and number of events (deaths) do not explain the different numbers of prognostic marker genes found for different cancers, and rather three other possible tumour specific parameters for these differences are suggested: intra-disease heterogeneity or responses to treatment that may act as putative confounders in the Cox model, and the possibility of differing levels of transcriptional dysregulation is also referred.

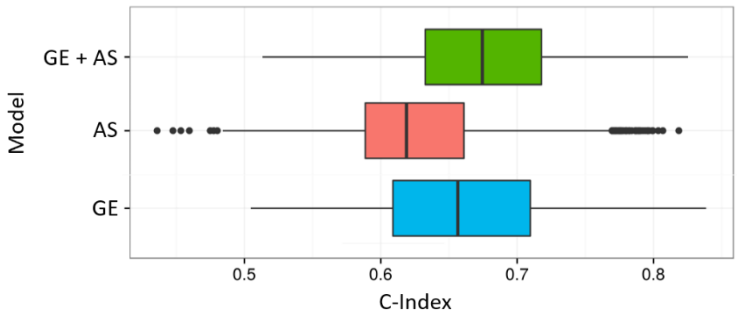


Figure 3.22 – Distribution of concordance indexes of Cox hazards-models for individual genes and alternative splicing events with prognostic value at Cox adjusted p-value below 0.01. GE – gene expression; AS – alternative splicing; C-Index – Concordance index.

Distribution of concordance indexes of Cox hazards-models for individual genes and alternative splicing events. GE – gene expression; AS – alternative splicing; C-Index – Concordance index.

We decided to further test if any of the alternative splicing events or expressed genes could add predictive power to the information already brought-in from DNA-methylation cluster, grade and age, as published in (Ceccarelli et al., 2016). This multivariate model was indeed confirmed to perform the best in the RNA-seq cohort used in the present study (in the published work, clinical

cases studied included both microarray and RNA-seq transcriptomic samples), presenting a concordance index of 0.867 (Figure S4). Then, multivariate Cox regression models were rerun for all genes and alternative splicing events, adjusting for DNA-methylation cluster, grade and age. Alternative splicing event models were also adjusted for the respective cognate gene expression, so as to be able to identify events with prognostic value independent of that of their cognate genes' overall transcript abundance.

From this analysis, at an FDR cut-off of 0.01, there were two alternative exons with prognostic value (Table 3.4): exon 2 of the *NLGN4X* gene, that codes for a neuroligin family protein, responsible for neuronal synapse remodelling, and exon 3 of gene *PDGFRA*, one of the RTKs that appears amplified in glioma at higher frequencies (2-4 %). However, these alternative splicing events had extremely low PSI variances to be considered regulated events (variances below 1 x 10^{-4}), with the NLGN4X-involving event having a hazardous effect with a mean PSI value of 0.0005 and zero interquartile range, and the PDGFRA-involving event being protective but with mean PSI value of 0.9968 and interquartile range of 0.0028. For this reason, they were considered not to have any biological interest and to be too difficult to use as prognostic markers. As for the equivalent Cox regression models for gene expression, at an FDR cut-off of 0.05 there was one gene that appeared to add prognostic value in glioma: *C1orf51* or *CIART*, a gene involved in circadian rhythm regulation in the mouse liver (Annayev et al., 2014) and the human prefrontal cortex (Chen et al., 2016). This gene showed a protective effect (HR 0.62) (Table 7).

Because the grade IV glioblastoma multiforme samples have markedly distinct gene expression and are suggested by PCA plots to be more homogeneous in relation to grades 2 and 3 ones, one could consider that some LGG-exclusive clinically relevant gene expression and splicing alterations failed to be detected in Cox models including the whole glioma cohort. Cox regression models for gene expression and alternative splicing, with adjustment for DNA-methylation cluster, grade and age were thus run again only with grades II and III patients, in order to detect yet some additional prognostic markers. At a 0.05 FDR threshold, two genes were detected: *C1orf51* and *TGIF1*, the latter a conserved transcription regulator that belongs to the TGFβ pathway (Table 3.3). As for the Cox models including alternative splicing events, there were eight events found to significantly improve performance (Table 3.3). Apart from exon 4 of gene *NHSL1*, whose PSI variance was of 0.015, all the others had again too low variances (below 1 x 10^{-4}) to be considered interesting as prognostic markers. *NHSL1* is a gene with unknown function and actually the one whose model produced the best concordance index (Table 3.3).

Table 3.3 – Cox proportional hazards models for prognostic maker genes, after adjustment for DNA-methylation cluster, grade and age. Number of events: 207 out of 627 patients. p-values shown are for the log-rank test and relative to the HR estimate of each variable, being all significantly against the null hypothesis of the variable not having an effect on survival (α = 0.001). HR – Hazards Ratio; GE – gene expression.

Model	Variable	HR	FDR	Concordance index
pan-Glioma: GE level + DNA-methylation cluster + Grade + Age	C1orf51	0.62	< 0.05	0.876
LGG: GE level + DNA-methylation cluster + Grade + Age	C1orf51	0.50	< 0.05	0.859
LGG: GE level + DNA-methylation cluster + Grade + Age	TGIF1	1.98	< 0.05	0.858

The number of prognostic markers found in this analysis was quite reduced, even after increasing the FDR cut-off for the Cox regression coefficient estimate from 0.01 to 0.05. Therefore, it may be concluded that alternative splicing and gene expression individual markers do not add further prognostic information for stratification of glioma patients.

Table 3.4 – Cox proportional hazards models for prognostic maker alternative splicing events, after adjustment for gene expression, DNA-methylation cluster, grade and age. Number of events: 207 out of 627 patients. p-values shown are for the log-rank test and relative to the HR estimate of each variable, being all significantly against the null hypothesis of the variable not having an effect on survival ($\alpha = 0.001$). HR – Hazards Ratio; GE – gene expression; AS – alternative splicing.

Model	Variable	HR	FDR	Concordance index
pan-Glioma: PSI + GE level + DNA-methylation cluster + Grade + Age	NLGN4X	2.8×10^{29}	< 0.01	0.871
pan-Glioma: PSI + GE level + DNA-methylation cluster + Grade + Age	PDGFRA	1.0×10^{-9}	< 0.01	0.875
LGG: PSI + GE level + DNA-methylation cluster + Grade + Age	CACHD1	0.00	< 0.05	0.836
LGG: PSI + GE level + DNA-methylation cluster + Grade + Age	FZD6	0.00	< 0.05	0.834
LGG: PSI + GE level + DNA-methylation cluster + Grade + Age	MEST	Inf	< 0.05	0.840
LGG: PSI + GE level + DNA-methylation cluster + Grade + Age	NLGN4X	7.0×10^{33}	< 0.05	0.847
LGG: PSI + GE level + DNA-methylation cluster + Grade + Age	NHSL1	1.1×10^{-2}	< 0.05	0.852
LGG: PSI + GE level + DNA-methylation cluster + Grade + Age	HKR1	7.9×10^{4}	< 0.05	0.845
LGG: PSI + GE level + DNA-methylation cluster + Grade + Age	RBM42	1.3×10^{-218}	< 0.05	0.845
LGG: PSI + GE level + DNA-methylation cluster + Grade + Age	IMMT	1.85×10^{45}	< 0.05	0.844

Next, we enquired about the existence of interesting, alternative splicing related, prognostic markers associated with the very important histological parameters tumour grade and molecular parameter DNA-methylation cluster. In particular, it was reasoned that prognostic markers associated with grade, independently of DNA-methylation cluster and age, and therefore with tumour progression, would be statistically significant variables in the models adjusted for DNA-methylation cluster and age. Similarly, and more interestingly, it was also thought that prognostic markers associated with

DNA-methylation cluster, independently of grade and age, would be statistically significant variables in models adjusted for grade and age.

New series of multivariate Cox proportional-hazards models were run for gene expression and PSI levels, in combination with, on the one hand, grade and age, and on the other, DNA-methylation cluster and age, whose results in terms of predictive value are shown in Figure 3.23A. At an FDR cut-off of 0.01, there were 3969 genes that produced improved Cox models in relation to grade and 3727 after further adjustment for age at diagnosis. At the same cut-off, there were 346 alternative splicing events adding prognostic value to the model adjusted for grade and 237 to the model adjusted for both grade and age (apart from the cognate gene expression adjustment). These affected 209 genes, 75 of which were not prognostic markers in terms of gene expression. These gene- and splicing event-including models have concordance indexes still below 0.867, the maximum for this cohort (*see* above), with the exception of models for genes *TGIF1, EMP3, TNFRSF12A* and *TIMP1*, and six including both expression and alternative splicing involving genes *BID, TNFRSF12A, PSTPIP1, TNK2, POLL*. In the latter though, alternative splicing events had non-significant Cox hazard ratio estimates at FDR 0.01 or 0.05. Nevertheless, the 237 alternative splicing events and 3727 genes appeared as a quite interesting set of prognostic markers to be analysed, namely by their putative association with particular LGm DNA-methylation clusters.

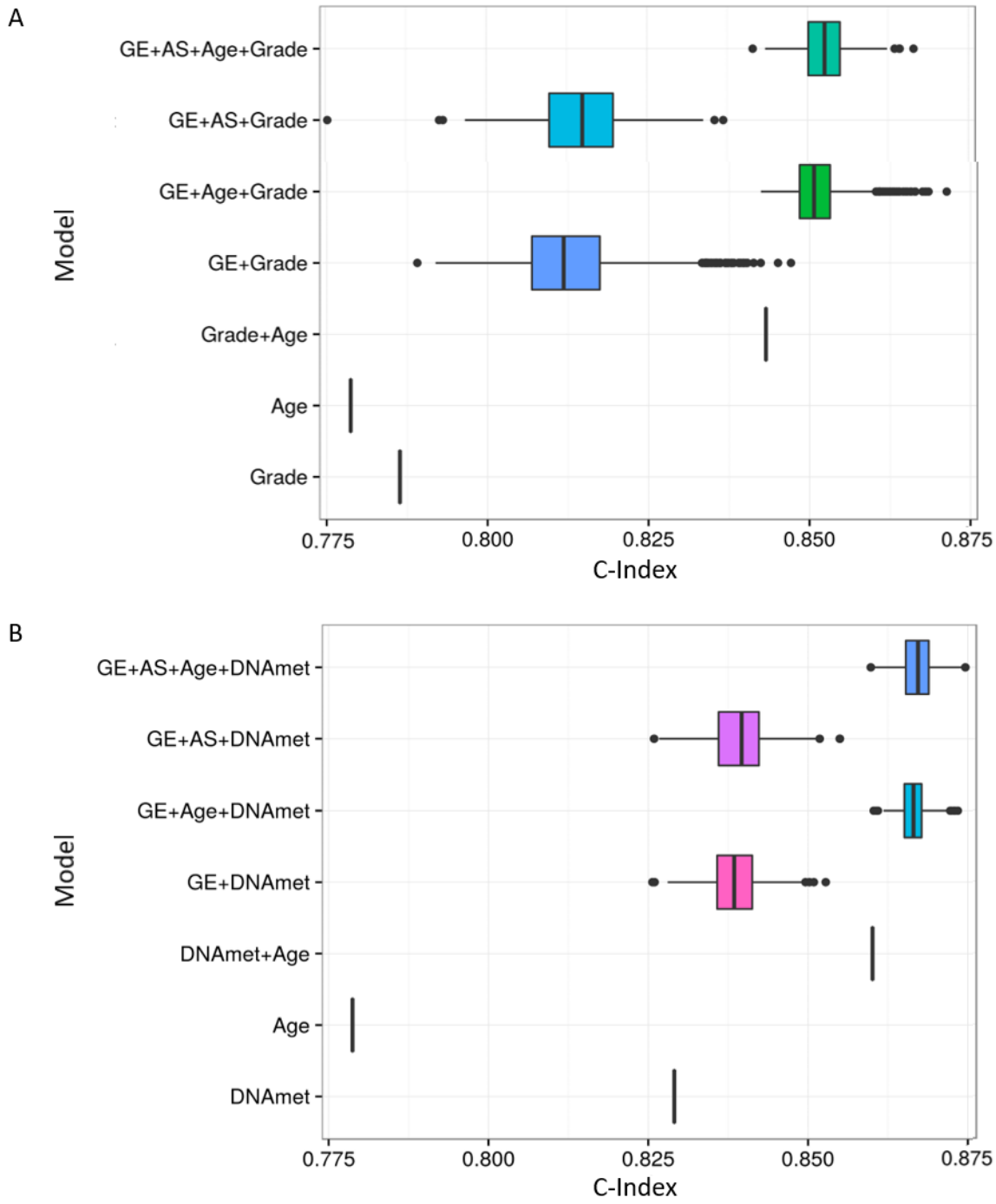


Figure 3.23 – Distribution of concordance indexes of Cox proportional-hazards models for individual genes and alternative splicing events with prognostic value at Cox adjusted p-value below 0.01. (A) Models adjusted for Age and/or Grade; (B) Models adjusted for Age and/or DNA methylation subtype GE – gene expression; AS – alternative splicing; C-Index – Concordance index; DNAmet - DNA methylation subtype.

Finally, as an attempt to discover additional prognostic alternative splicing events after adjustment for grade, age and gene expression, Cox regression models run only for the LGG cohort were generated. This resulted in the discovery of 675 events at an FDR cut-off of 0.01, affecting 557 genes, 244 of which were not prognostic markers. The 493 novel prognostic alternative splicing events detected in the LGG cohort were also gathered for analysis of association with the LGm groups.

There were 553 genes that produced improved Cox models in relation to DNA-methylation cluster and 130 after further adjustment for age at diagnosis (Figure 3.23B). Models including alternative

splicing yielded 63 events adding prognostic value (FDR < 0.01), after adjusting for DNA-methylation cluster, and, after adjusting for both DNA-methylation cluster and age, the same two previously found for the models including grade (Table 3.4). The fact that the same alternative splicing prognostic markers were found for Cox models that included or excluded adjustment for grade suggests that alternative splicing regulation does not contribute, at least in great extent and in a way that is independent from gene expression, to grade-associated glioma prognosis prediction after adjustment for DNA-methylation cluster. Still, it should be added that, in the models that include DNA-methylation and age but not grade, if an FDR cut-off of 0.05 is in turn used, there are additional 39 alternative splicing events that become significant contributors to the multivariate Cox model, which may still be looked-upon as good candidates to be grade and thus malignancy degree-associated markers in glioma. Statistics for the Cox regression models including these 41 alternative splicing events are presented in Table S2.

In order to focus on alternative splicing related prognostic markers potentially associated with glioma DNA-methylation cluster, only the genes and alternative splicing events that showed to be able to add prognostic value to glioma patient overall survival prediction in addition to grade and age will be considered from this point on.

3.2.3 Identification of potential *trans*-acting regulators of splicing in different DNA-methylation subtypes

In order to discover likely mechanisms of regulation in *trans* of alternative splicing in glioma DNA-methylation subtypes, genes coding for RNA-binding proteins (RBPs) and splicing factors (SFs), namely RNA-binding ones, were identified among prognostic markers and differentially expressed genes described in the previous sections.

Among the 3727 good prognostic gene markers after adjustment for grade and age, there were 328 RNA-binding proteins and 75 splicing factors from the list taken from (Sebestyén, Singh, et al., 2015). From these, 106 RBPs and 20 SFs were differentially expressed across DNA-methylation clusters, at FDR < 001 and fold change > 2 (Figure 3.24A).

Focusing on SFs with a known RNA-binding motif would enable an *in silico* search for their putative pre-mRNA targets. 17 RBP and SF prognostic markers had known RNA-binding motifs, six among the differentially expressed across DNA-methylation subtypes: IGF2BP2, IGF2BP3, KHDRBS2, PCBP3, RBM47 and RBMS1 (Figure 3.24B). These RBPs were thus selected for further investigation on the mechanisms of alternative splicing regulation in glioma.

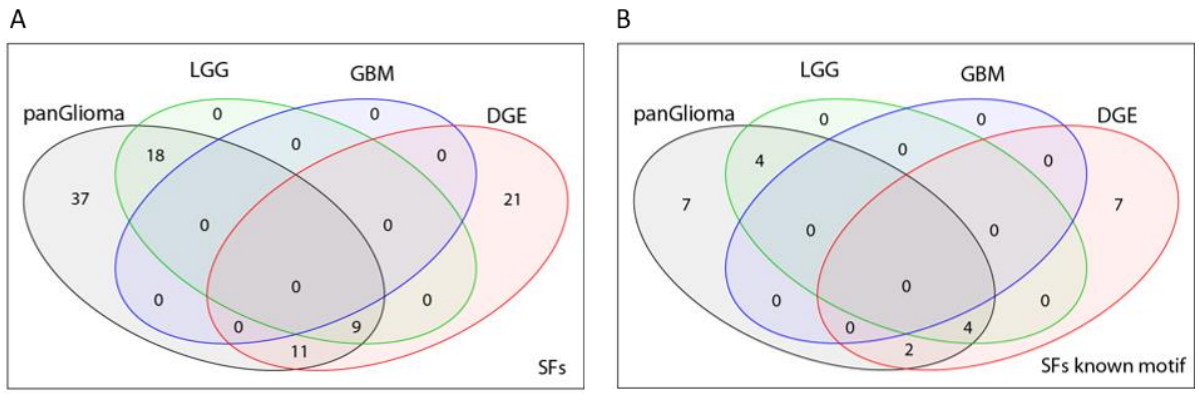


Figure 3.24 – Prognostic splicing factors associated with LGm subtype. Numbers are relative to genes whose predicting adjusted p-values in multivariate Cox regression models adjusted for tumour grade and patient age at diagnosis were below 0.01, except for the DGE sets, which refer to differentially expressed splicing factor genes across LGm subtypes. (A) Venn diagram for splicing factor genes (SFs); (B) Venn diagram for splicing factor genes coding for proteins with known RNA-binding motif.

Before moving into the next section, a note may be left about other promising RNA-binding splicing factors that will not be included in the study of regulation of splicing in *trans* but that are still promising alternative splicing regulators for having a relevant function in glioma. Splicing factors like A2BP1 (or RBFOX1), PABPC5, RBM24, RBM42, YBX1, CELF4 and CELF5, although having been found to work as good glioma prognostic markers, ceased to contribute to prognosis prediction when glioma grade and patient's age were accounted for. On the other hand, genes like PABC1, HNRNPA1L2, TUT1, HNRNPA1, FXR2 and KHDRBS3, which still added prognostic value to Cox regression models with grade and age adjustments, did not meet the cut-off value of fold changes between at least two DNA-methylation groups to be included in the group of differentially expressed genes. Still these SFs had log2-fold change differences of around 0.9 between at least one pair of LGm subtypes. Because the search for mechanisms of splicing regulation relied on the detection of splice ratio switches between groups of samples that also differed in their expression of each candidate splicing regulator, it seems important to guarantee that the magnitude of these expression changes is high enough.

3.2.4 Identification of DNA-methylation subtype associated prognostic alternative splicing events

From the 237 alternative splicing event prognostic markers identified, 122 were in fact differentially spliced between DNA methylation subtypes at the criteria described in section 3.1.6. From the 675 identified after survival analysis performed only on the LGG cohort, an additional 215 appeared differentially expressed (Figure 3.25).

From these 337 events with significant prognostic value and differentially spliced across LGm subtypes, 45 were on 40 differentially expressed genes (FDR < 001 and log2 fold-change > 1) (Figure 3.25B). There were 246 alternative splicing events showing statistically significant correlation with their gene expression, affecting a total of 208 genes. Although these alternative splicing events might be affected by one of the putative RBPs specific of DNA-methylation clusters, this other dependence on own gene expression might interfere with the detection of such regulatory function. In addition, 221 (affecting 193 genes) out of the 337 AS events of interest already corresponded to genes detected as prognostic markers. There were 50 events that appeared as gene-expression independent prognostic markers (Figure 3.25), mostly skipped exons and alternative-first exons.

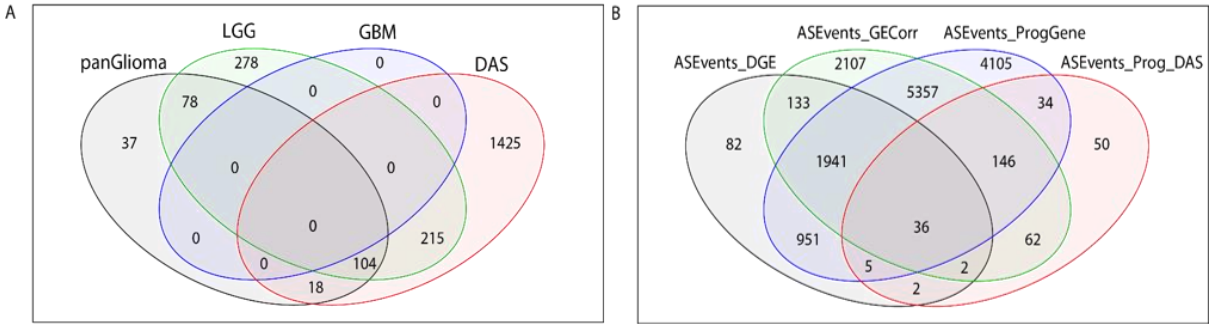


Figure 3.25 - Relations between alternative splicing prognostic markers and alternatively spliced and differentially expressed genes. (A) Venn diagram for the intersection of different sets of alternative splicing events with predictive adjusted p-values in multivariate Cox regression models adjusted for grade (except for GBM models) and age below 0.01. Numbers were obtained from Cox regression models applied to all glioma (panGlioma), only low-grade-glioma (LGG) and only glioblastoma (GBM) samples. DAS - differentially alternatively spliced events across LGm subtypes. (B) Venn diagram for the intersection of different sets of alternative splicing events: affecting genes that are differentially expressed (ASEvents_DGE), whose PSIs correlate with the expression of cognate gene (ASEvents_GECorr), which affect genes that have prognostic value independent from grade and age (ASEvents_ProgGene), which have prognostic value independent from grade and age (ASEvents_ProgDAS).

Finally, PCA plots for the 337 alternative splicing prognostic markers whose association with LGm groups was determined here are shown in Figure 3.25. These show the ability of this selected pool of events to perform a refined distinction between epigenetic subtypes.

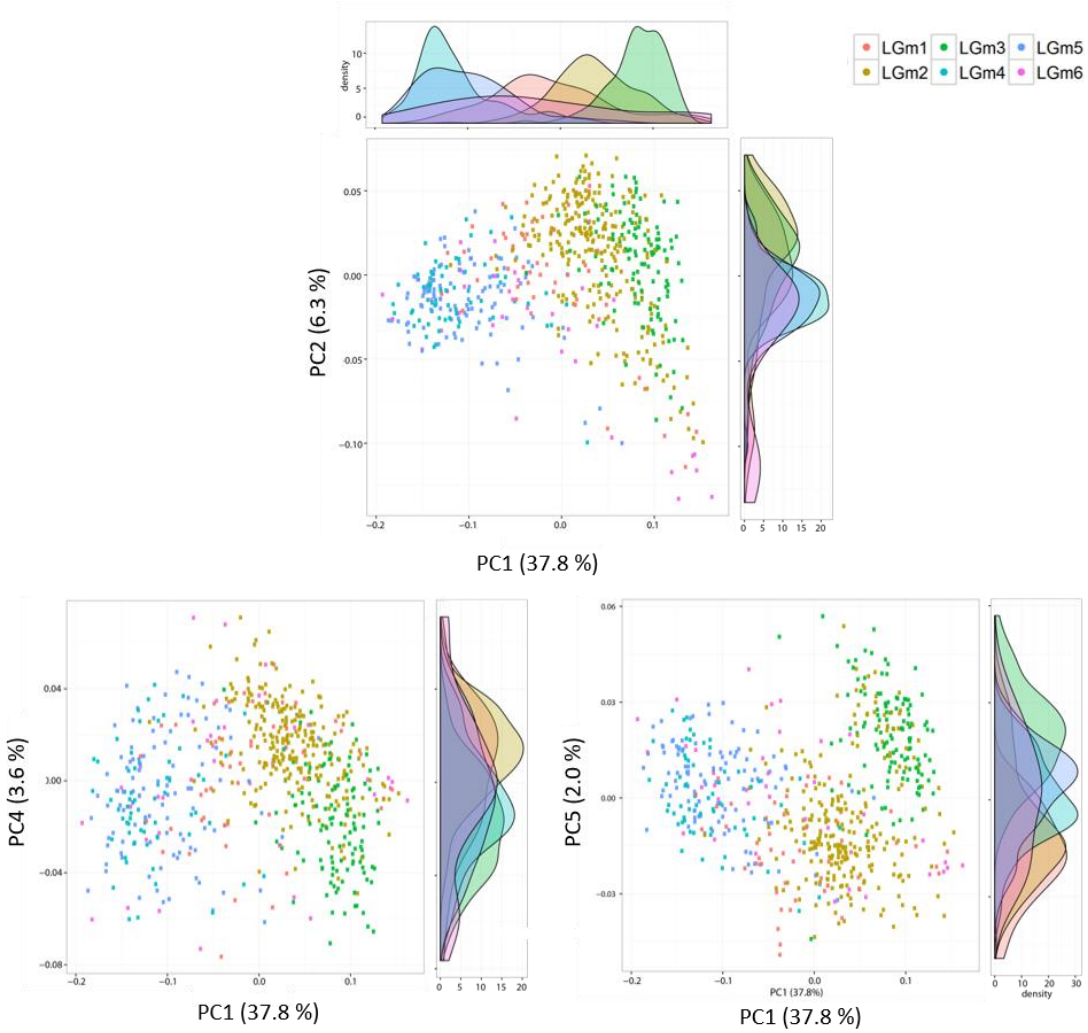


Figure 3.26 – Principal Component Analysis plots made on 337 prognostic AS events associated with LGm subtypes. Selected PCs are plotted. Colours are according to DNA-methylation subtype.

3.3 DISCOVERY OF ALTERNATIVE SPLICING REGULATION MECHANISMS IN GLIOMA

In recent years, great advances have been made in terms of the identification of splicing regulatory elements and factors, among which RNA-binding proteins. To investigate alternative splicing regulatory mechanisms in *trans* in a particular tissue of a model organism or cell line, one can combine alternative splicing profiling in a control and loss-of-function *in vivo* model for a given splicing factor with CLIP-seq technology that will allow to identify mRNA-binding regulatory regions for the same splicing factor, and in turn produce a very good set of potential targets of splicing regulation. However, while not all RNA-binding splicing factors have been subject to this kind of detailed studies, *in silico* approaches can help generating this very same kind of hypotheses. Tools like the FIMO software (Grant, Bailey, & Noble, 2011) allow to, from the knowledge acquired *in vitro* about RNA-binding protein (RBP) binding preferences, map motifs along the genome and subsequently identify regions of higher density of motifs in the vicinity of regulated exons. Information taken from the application of this kind of analysis is able to compensate for the lack of CLIP-sequencing availability.

From the occurrence of an RBP binding motif in intronic or exonic regions flanking a regulated exon, one can infer a good candidate alternative splicing regulatory region. However, it is not possible to guess if it will function as an enhancer (ESE/ISE) or as a silencer (ESS/ISS). But, if one considers that the level of expression of an RBP splicing factor will correlate with the levels of inclusion of its target exons, then the enhancer *vs* silencer nature of the RBP on those targets may be extracted from that correlation. Recently, the Genotype-Tissue Expression (GTEx) project made available a set of RNA-seq data from thousands of *post-mortem* samples from 32 different tissues coming from healthy individuals (Lonsdale et al., 2013). The idea of combining the transcriptomic quantification from GTEx with the running of FIMO with known RBP-binding motifs, in order to build RNA binding maps for splicing regulators, arose in the lab.

In this section, our attempt to implement this strategy for RBP splicing regulatory mechanism discovery will be described. GTEx was used as a powerful large data set coming from normal-functioning tissues that allows the best possible outline of each RNA splicing map. The same approach was then applied to the TCGA data set. Specifically, a focus will be put on trying to relate prognostic splicing factors with prognostic alternative splicing events differentially regulated in LGm groups.

3.3.1 On the likeliness of glioma prognostic alternative splicing being mediated in *trans*

From the previous section, among the prognostic markers that appeared strongly associated with DNA-methylation subtype differences, there were 337 alternative splicing events and six RNA-binding splicing factors with a known binding motif.

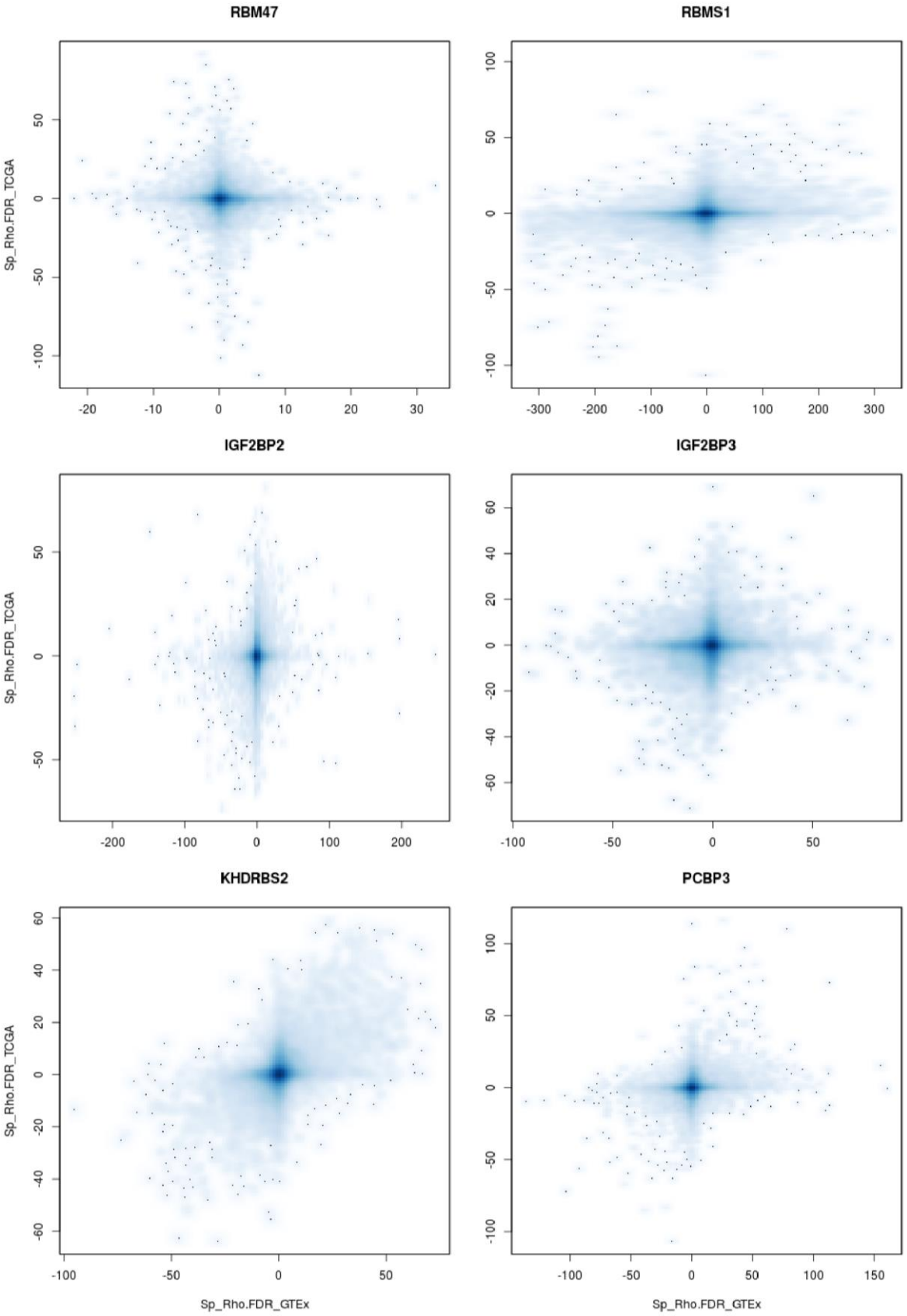


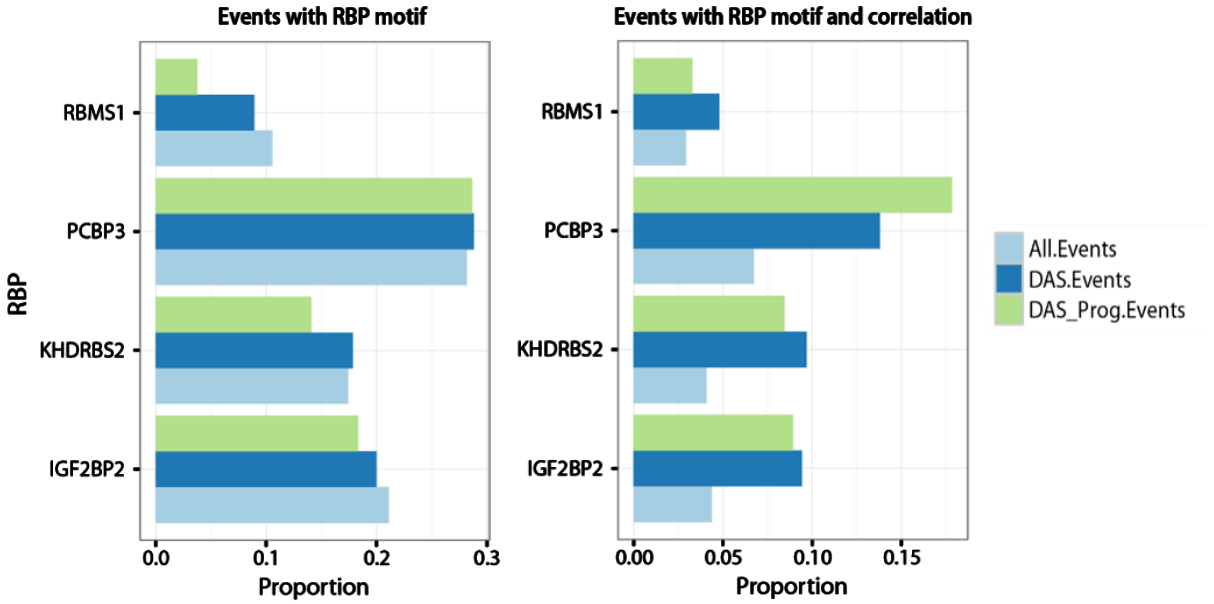
Figure 3.27 – Concordance between glioma TCGA and GTEx splicing factor expression to alternative splicing events PSIs correlations. Scatter plots of -log10(FDR) values taking positive or negative values according to the sign of the rho of Spearman.

To ascertain the odds for each of the six RBPs to have a significant role in determining alternative splicing regulation decisions, we inspected whether RBP expression levels concordantly (i.e. with the same positive or negative sign) correlated in the TCGA glioma and the GTEx cohorts. Scatter plots relating the adjusted p-values of tests for correlation between RBP expression and PSIs were produced (Figure 3.27). While for RBM47 and IGF2BP3 no trend for an agreement in sign and strength of correlations was detectable, this was more the case for the remaining proteins, especially for KHDRBS2, which encodes a protein highly expressed in the brain (data not shown). To have a comparison of how this same kind of plot would look like for RBPs known to have preponderant effects in alternative splicing regulation, PTBP1 and A2BP1 (RBFOX1) correlation test results were also plotted (Figure S5).

The hypothesis that the four RBPs that showing correlation with exon inclusion ratios similar in the cancer and normal tissue samples in fact regulate some of the alternative splicing events of interest through direct binding was evaluated.

Frequencies of binding motifs in splicing regulatory sequences were checked for, without and with the additional requirement of significant correlations between RBP gene expression and PSI in both GTEx and TCGA samples (Figure 3.28).

Indeed, there were binding motifs for each of the RBPs in the groups of differentially spliced events (Figure 3.28, left panel). In addition, for all *RBMS1, PCBP3, KHDRBS2* and *IGF2BP2* factors, there is a larger proportion of events having RBP motif together with significant correlation with RBP expression among differentially spliced events when compared to the background of all alternative splicing events (Figure 3.28, right panel).



	With RBP motif			*With RBP motif and RBP correlation*		
	All events (N=8894)	***DAS events (N=1166)***	***DAS&Prog events (N=213)***	***All events (N=8894)***	***DAS events (N=1166)***	***DAS&Prog events (N=213)***
RBMS1	1129	104	8	313	56	7
PCBP3	3017	336	61	2239	528	130
KHDRBS2	1867	208	30	437	113	18
IGF2BP2	2257	233	39	467	110	19

Figure 3.28 – Evidence for alternative splicing regulation by four RBPs. Barplots show the proportions of all 17151 events ("All.Events"), differentially spliced events ("DAS.Events") and differentially spliced events that have prognostic value

independently of tumour grade and patient's age ("DAS_Prog.Events") that meet two different criteria: on the left, presence of RBP-binding motif; on the right, presence of RBP-binding motif and correlation with RBP expression (Spearman FDR < 0.01) in both glioma TCGA and GTEx samples. The table shows the same information in absolute frequencies.

3.3.2 RNA splicing maps

Here, the regulation of the most frequent event type, exon skipping, will be studied. An alternative exon located between two constitutive exons will be spliced as efficiently as its exon-defining 3'- and 5'-splice sites are recognized by the spliceosome.

In this section, we attempt to discover whether any of KHDRBS2, PCBP3, IGF2BP2 or RBMS1 proteins has a mode of regulation that obeys to defined spatial rules, i.e. their splicing enhancing or silencing roles are determined by the distance of their binding to mRNA to nearby splicing sites.

Essentially, different combinations of thresholds of significance for, on the one hand, correlations between PSIs and RBP expression levels and, on the other, for FIMO p-values for each given RBP binding motif were used in order to find which set of these cut-off parameters maximizes the strength of association between RBP-expression level-dependent splicing ratios and presence of RBP-binding motifs. The strength of association was measured with the one-sided Fisher's Exact test, taking the alternative hypothesis that there is a higher proportion of alternative splicing events correlated with RBP gene expression when the events' regulatory regions are enriched for RBP binding motifs. Tests for positive (enhancing of exon inclusion) and negative (silencing of exon inclusion) correlations were made separately for each of eight regulatory regions adjacent to the splice sites of the implicated alternative and constitutive exons (Figure 3.29, *see* Methods), for each combination of cut-off parameters. The cut-off parameters for the two regions that returned the best result on the Fisher's exact test were then used to create RNA splicing maps, which result from the application of the same statistical test along the 800 bp of regulatory regions defined, using a sliding window spanning 50bp (*see* Methods for a detailed explanation).

This approach was validated using PTBP1, the ubiquitous splicing factor known to silence exon inclusion through binding to the intronic region that is right upstream of it and to enhance exon inclusion when binding the immediately downstream intronic region (Raj & Blencowe, 2015). The results are presented in Figure S6 and show clearly a peak of silencing regulation at around 40 bp upstream of the regulated exon, remarkably visible both for the GTEx normal tissue (Figure S6A) and the glioma TCGA (Figure S6B) datasets. In addition, the effect of silencing appears along the whole 150 intronic region upstream from the regulated exon (region s2_I) and also in the last 50 bp of the same exon (region e2_E). As for the known PTBP1 enhancing effect upon binding to the downstream intron, this appears for one of the "GTEx" maps and also for one of the "TCGA" maps. These results effectively validate the power of the approach presented here for the discovery of alternative splicing regulation rules. The fact that results were in great part concordant for the two data sets indicates that, first of all and as already known from its described role in supporting glioblastoma progression, PTBP1 is functional in the glioma tissue and thus that this methodology can be used to find robust hypotheses for mechanisms of alternative splicing regulation acting in cancer tissue. The fact of having less significant results for the TCGA cohort relates at least in part to the fact that the number of alternative splicing events used to generate these maps is much lower: 4859 alternative splicing events as compared to 14769 from GTEx.

Next, RNA splicing maps were generated for the other four splicing factors of interest in this glioma study. While using the GTEx data set to find regulatory regions for RBMS1, the most significant Fisher's p-value obtained for any of the regions and threshold parameters was above 0.01,quite high

in comparison to what was obtained for the other genes. So, the analysis of RBMS1 alternative splicing regulatory role was dropped at this point. It is possible that this protein's role on alternative splicing does not follow this direct mode of regulation.

As for PCBP3, it is expressed at higher levels in the brain, the cerebellum, the pituitary and the testis (data not shown) and its expression has been shown to be particular of post-mitotic, differentiated cells, similarly to what happens with A2BP1. It showed as most promising alternative splicing regulatory regions the intronic mRNA segment that is located upstream from the second constitutive exon (s3_I) and the intronic region that follows the alternative exon (e2_I) (Figure 3.29). In both cases, the result of the Fisher's test were significant for enrichment of binding motifs in events that correlate negatively with *PCBP3* expression. When plotting RNA maps using the FIMO and correlation FDR cut-offs that returned these stronger enrichments, two peaks representing PCBP3 binding regions resulting in alternative exon silencing were commonly observed in the two maps. However, there were also peaks that were different, namely the enhancing region at around position 50 of region s2_I and the silencer region just before the start position of the second constitutive exon in region s3_I. It is possible that PCBP3 has a role in favouring exon skipping in events whose regions e2_I have very high affinity binding motifs (corresponding to 3.83 FIMO p-value threshold of the second map) and either an alternative exon splicing role or exon skipping for event whose regions s2_I or s3_I have lower affinity binding motifs (corresponding to the less stringent 3.05 FIMO p-value threshold of the first map).

RNA splicing maps for PCBP3 using the TCGA dataset looked unrelated with the ones previously seen for the GTEx dataset, whatever the FIMO p-value thresholds used. Regions where there was a higher enrichment of PSI to PCBP3 expression correlation upon putative binding of PCBP3 were exonic regions e2_E and e1_E, which indeed presented enhancing peaks in each of the two PCBP3 splicing maps. It remains to be ascertained if in glioma, changes in factors other than PCBP3 expression, like PCBP3 protein turnover or activity, or else alterations in its protein interactors, might be causing these different responses to PCBP3 transcriptional output.

Because PCBP3 presented very strong correlations between its expression and a relatively (in comparison with KHDRBS2, PTBP1 or A2BP1) low number of alternative splicing events, both in the GTEx and the TCGA cohorts, it may be responsible for the regulation of a minority of alternative splicing events due to the requirement for stronger binding affinity. This would favour the e2_I map of GTEx as the most reliable (the second map in Figure 3.29A), whose best parameters for a splicing RNA-binding plot actually were the same as for the ones of another region: the end of the first constitutive exon.

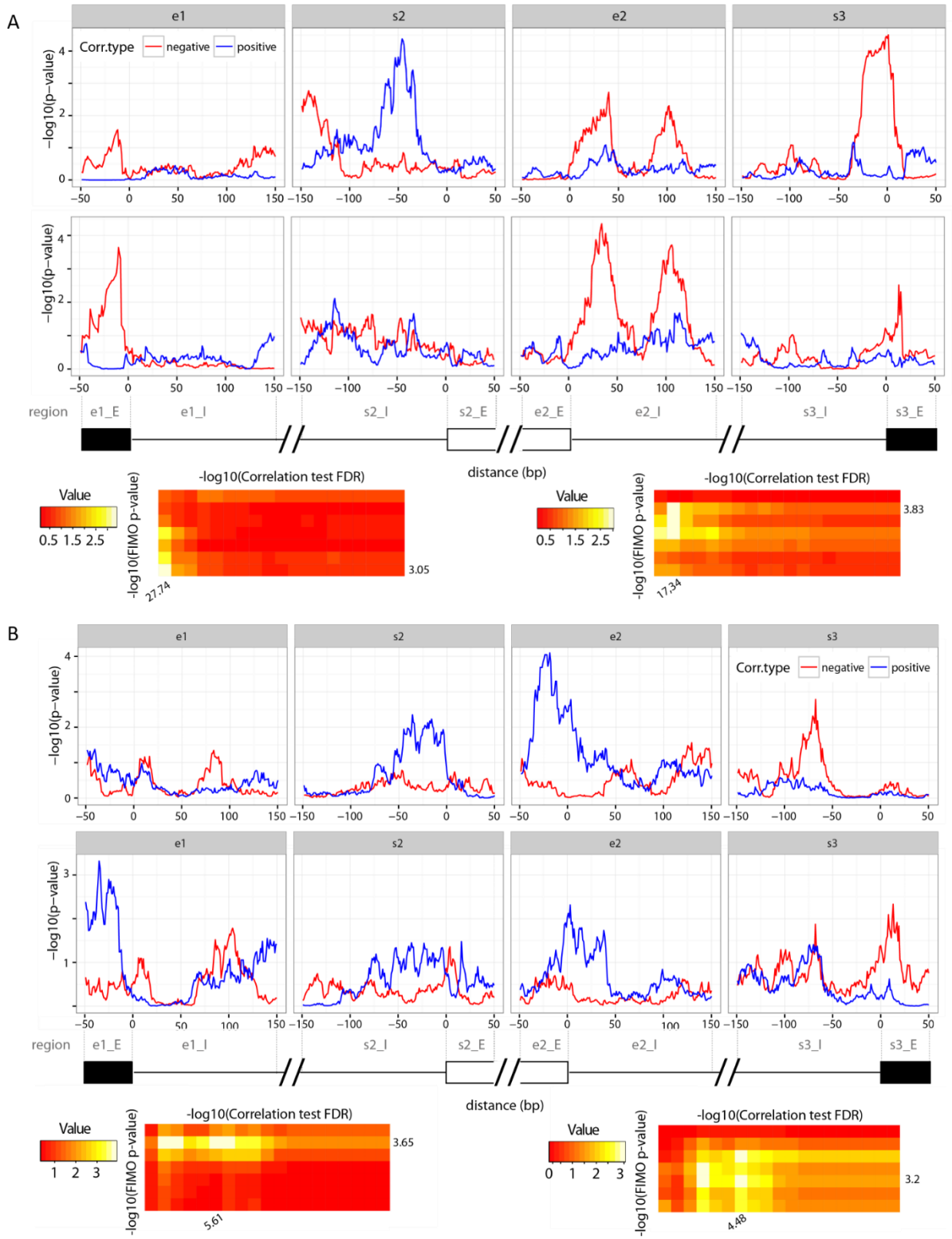


Figure 3.29 – PCBP3 RNA-binding maps for the general exon-skipping (SE) alternative splicing event. (A) Two RNA-binding maps produced using the GTEx multi-tissue dataset. (B) Two RNA-binding maps produced using the glioma TCGA dataset. RNA-binding maps shown were generated using correlation FDR threshold and FIMO p-value as shown on the bottom and the right side of the heat maps, respectively. Distance in base pairs (bp) relative to the closest splice site is shown. Different names for the eight intronic (150 nucleotides long) and exonic (50 nucleotides long) regulatory regions defined are indicated in grey. Constitutive exons are shown in black and alternative exon is shown in white. Corr-type – Correlation type: blue for enhancement and red for silencing of exon inclusion.

KHDRBS2 is expressed mostly in the brain, thyroid, lung, pituitary, intestine and spleen (data not shown). It was the splicing factor that showed the highest proportion of alternative splicing events having concordant correlation test results for RBP expression *vs* events PSIs between the GTEx and the TCGA datasets (Figure 3.30).

RNA splicing maps for the GTEx datasets presented a peak of silencing activity for KHDRBS2 in the 50 bp of intronic region that precedes the alternative exon (Figure 3.30). This peak exhibited robustness to varying combinations of FIMO and correlation parameter thresholds (data not shown). The first and second maps were produced using cut-off values that maximized the Fisher's exact test for non-random association between negatively correlated RBP expression and PSIs and the presence of RBP binding motif for regions s2_I and e2_E, respectively.

Similar to what had been observed with *PCBP3*, RNA splicing maps derived from TCGA data looked very different from the ones obtained from the GTEx data. This was somehow unexpected because of the highly concordant correlation results referred above between the two datasets, which automatically implied that a similar pool of alternative splicing events with the same KHDRBS2 binding motifs in their mRNA regulatory regions would be accounted for during map generation. However, the amount of alternative splicing events used for the generation of the maps was much lower for TCGA than for GTEx: 4369 and 14769, respectively. This tremendous difference derives mostly from the fact that the TCGA genome annotation contemplated 17533 SE events, while the annotation used for analysis of GTEx data, the GENCODE annotation, included 35465 SE events. This difference is indeed illustrated by the difference in the maximum enrichment significance achieved with both datasets ($p \approx 10^{-2}$ vs $p \approx 10^{-6}$). Such weakness could be solved through the analysis of raw RNA-seq data for TCGA glioma cohort using the GENCODE genome annotation.

Finally, RNA splicing maps were also produced for IGF2BP2, a ubiquitous protein that appeared upregulated in LGm groups 1,4,5 and 6 in relation to LGm2 and LGm3 (Figure 3.30). In the case of this RBP, the two regulatory regions that showed better non-random association between the effects of RBP expression and the presence of RBP binding motif on splicing ratios were the same for the two data sets: the intronic region upstream from the regulated exon (s2_I) and the beginning of the second constitutive exon (s3_E), although statistical significance was higher for s2_I region in the GTEx dataset and for s3_E region in the TCGA dataset. Interestingly, for the GTEx data set no peaks for the s3_E region appeared in either map. Instead, apart from the silencer peak in s2_I region at around 90 bp from the start of the alternative exon, there was an enhancing peak at 50 bp of the same intronic region and also another enhancing peak spanning a large portion of region e2_I. These two peaks representing an enhancer activity of IGF2BP2 had been detected through the application of Fisher's exact test for positively correlated events on the 150 bp-spanning regions, though their corresponding p-values were not among the two highest. In the second TCGA RNA splicing plots, it is possible to observe a silencing and an activating peak in the intronic region that precedes the alternative exon, as much as in the GTEx maps, though at low levels of significance. However, in this map there are other peaks that could indicate an IGF2BP2 regulatory role, all of them having a low y-axis coordinate, that corresponds to a p-value above 0.01. The top map obtained from the TCGA cohort shows an enhancing peak of higher significance at exonic region s3_E, which was expected to appear for the GTEx cohort as well.

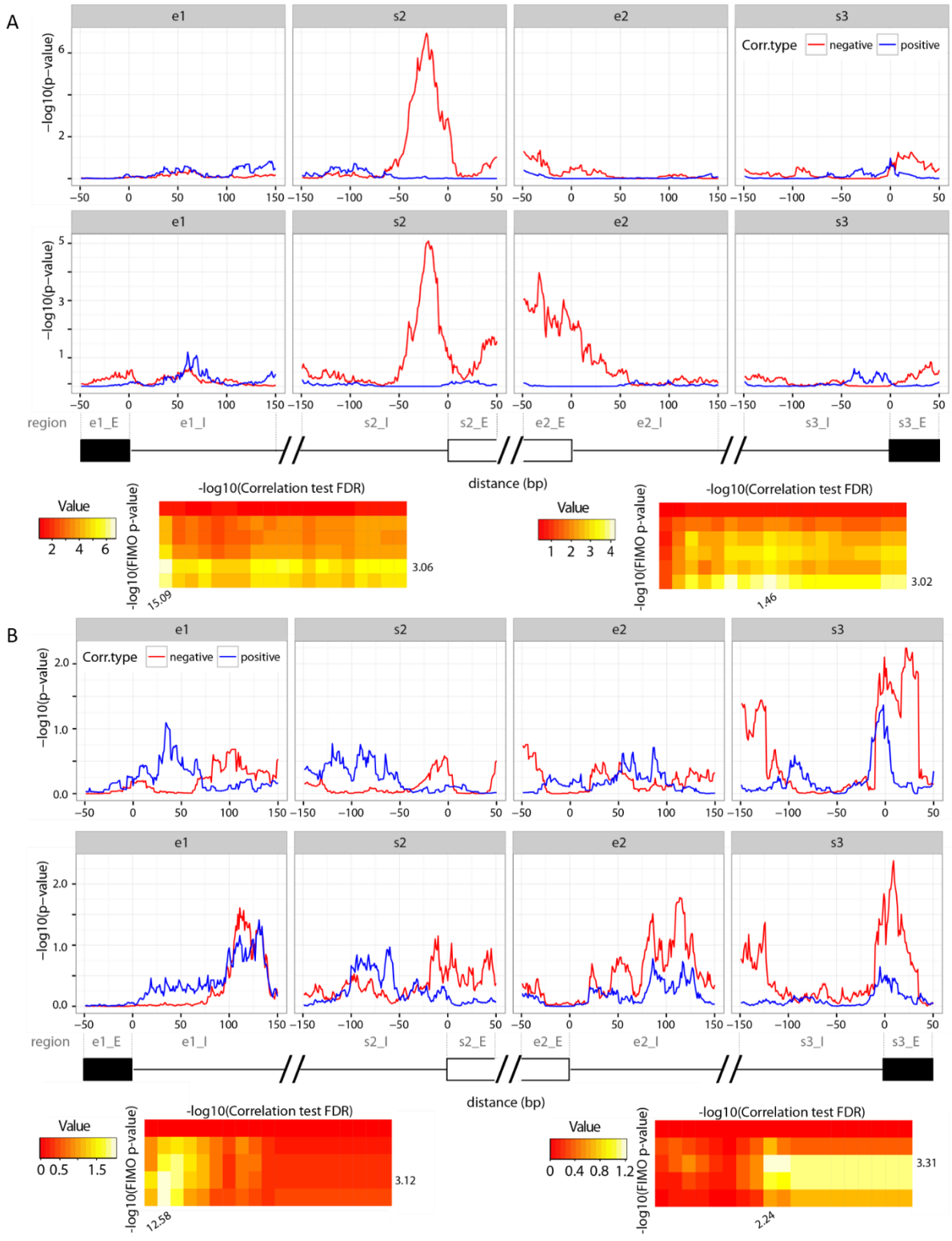


Figure 3.30 – KHDRBS2 RNA-binding maps for the general exon-skipping (SE) alternative splicing event. (A) Two RNA-binding maps produced using the GTEx multi-tissue dataset. (B) Two RNA-binding maps produced using the glioma TCGA dataset. RNA-binding maps shown were generated using correlation FDR threshold and FIMO p-value as shown on the bottom and the right side of the heat maps, respectively. Distance in base pairs (bp) relative to the closest splice site is shown. Different names for the eight intronic (150 nucleotides long) and exonic (50 nucleotides long) regulatory regions defined are indicated in grey. Constitutive exons are shown in black and alternative exon is shown in white. Corr-type – Correlation type: blue for enhancement and red for silencing of exon inclusion.

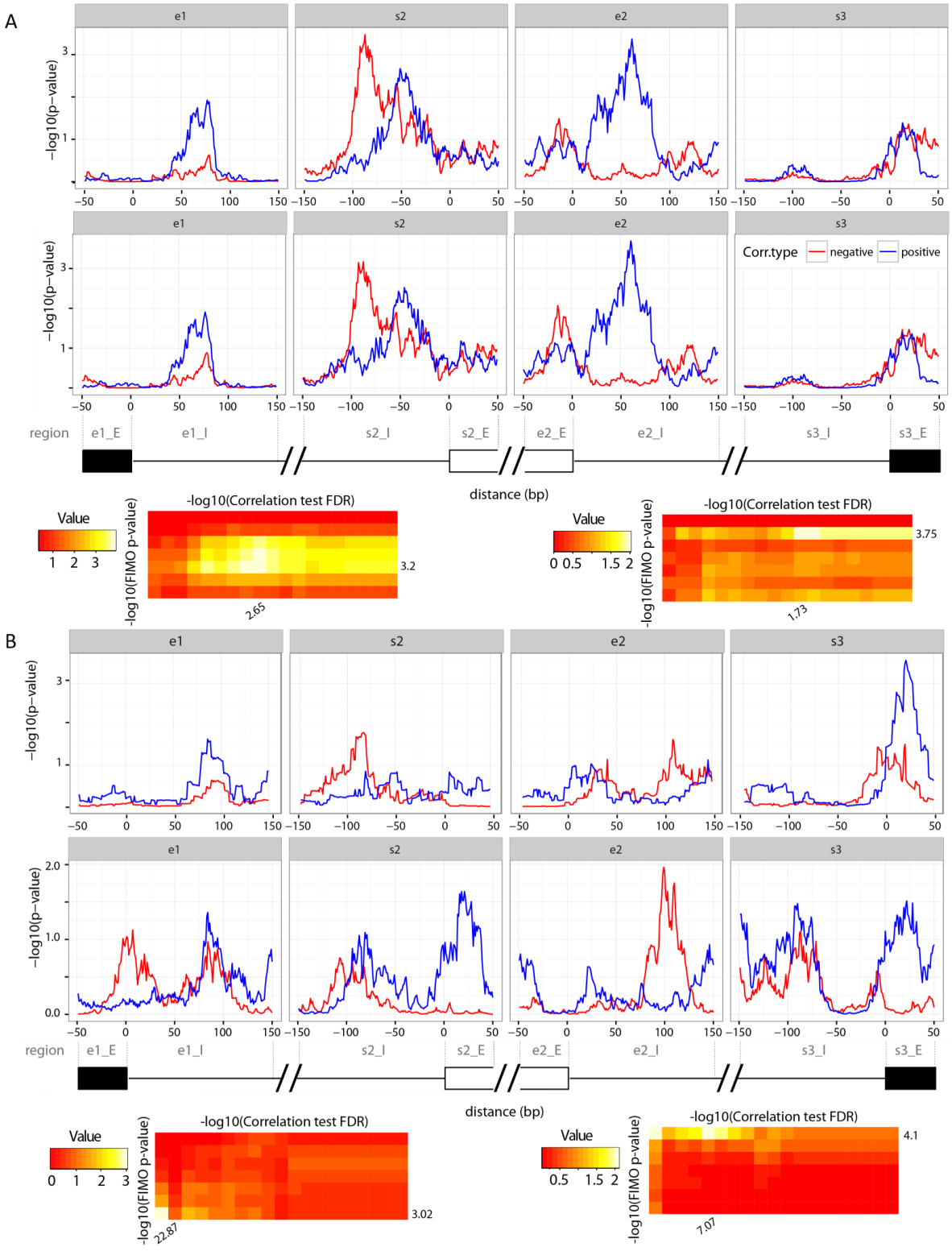


Figure 3.31 – IGF2BP2 RNA-binding maps for the general skipped exon (SE) alternative splicing event. (A) Two RNA-binding maps produced using the GTEx multi-tissue dataset. (B) Two RNA-binding maps produced using the glioma TCGA dataset. RNA-binding maps shown were generated using correlation FDR threshold and FIMO p-value as shown on the bottom and the right side of the heat maps, respectively. Distance in base pairs (bp) relative to the closest splice site is shown. Different names for the eight intronic (150 nucleotides long) and exonic (50 nucleotides long) regulatory regions defined are indicated in grey. Constitutive exons are shown in black and alternative exon is shown in white. Corr-type – Correlation type: blue for enhancement and red for silencing of exon inclusion.

In this section, the analysis of RNA splicing maps for five splicing factors showed the value of this novel approach of integrating alternative splicing profiles with RBP binding predictions in order to outline possible mechanisms of alternative splicing regulation. It also showed how the application of this procedure to two datasets can assist in forming more robust hypotheses that can then be tested experimentally. However, the use of much fewer alternative splicing events in the TCGA data may actually be the main cause for the incoherencies found in the RNA splicing map analyses.

4 DISCUSSION

To date there are two publications dedicated to the study of malignant glioma in a transversal way across grades 2 to 4 (Ceccarelli et al., 2016; Z.-L. Z. Wang et al., 2015), both dealing with GBM and LGG TCGA glioma cohorts. The study from Wang and collaborators allowed to identify a group of 1091 genes which at both mRNA and DNA-methylation level distinguished glioma samples in between grades and thus across levels of malignancy. Among the 977 genes showing upregulation in higher grades, the main enriched functional category was cell cycle. The work from Ceccarelli and collaborators was also a multi-platform integrative study in which the authors were able to create a glioma classifier based on 932 DNA-methylation probes, which has the resolution to distinguish more than the three previously identified glioma epigenetic classes: *IDH*-mutant non-codel, *IDH*-mutant codel and *IDH*-wiltype glioma classes. Indeed, *IDH*-mutant non-codel glioma was further divided in two with differing levels of DNA-methylation levels, with LGm1 subtype, the one with lower DNA-methylation, constituting a poorer prognosis group shown to frequently correspond to advanced stages of the disease in relation to the closely-related LGm2 subtype. Then, within *IDH*-wild type glioma, three subtypes could also be distinguished, with LGm6 constituting a heterogeneous, overall better prognosis class, sharing epigenetic and genetic characteristics with the often benign tumour pilocytic astrocytoma. The ability of this molecular classifier to diagnose novel groups with differential prognosis rendered it a quite powerful means of, together with the established glioma prognosis indicators grade and age, performing an evaluation of a patient's predicted disease outcome as well as investigating therapeutic strategies favourable to each particular molecular subtype, both dealing with GBM and LGG TCGA cohorts(Ceccarelli et al., 2016; Z.-L. Z. Wang et al., 2015).

This thesis focused in the analysis of the RNA-seq GBM and LGG data sets from the TCGA portal in order to establish the contribution of alternative splicing regulation to the definition of subtypes of glioma grades 2 to 4.

Initial exploratory data analysis of the 659 glioma cases cohort allowed to determine that the majority of alternative splicing events had their exon-inclusion ratios correlated with levels of transcription of their cognate gene. Evaluation of the association between gene expression and alternative splicing was carried out with the main purpose of identifying which events of splicing might be influenced by rates of transcription. This information could then be taken into account when studying mechanisms of alternative splicing regulation *in trans*, as the identified alternative splicing events would be known to have splicing ratios determined by an interaction between levels of gene expression and the actual relative abundances of active splicing factors. The set of alternative splicing events found to be more strongly associated with gene expression had a distribution of PSI variances higher than the set that displayed a weaker association. This difference could stem from the fact that some low variance events actually associated with gene expression could have failed to be detected by correlation analysis, which would have resulted in a reduced but biased number of low variance events among the set of events with high association with gene expression. This was interpreted as an indication that the group of events less associated with gene expression could be less rich in interesting alternative splicing events descriptive about differences between glioma grades and DNA methylation subtypes. However, while performing the analysis of differential splicing across LGm subtypes, it became clear that the ability for alternative splicing events to help distinguish glioma subtypes was frequently not dictated by the overall amount of variation of their PSIs.

Analysis of the main principal components of variance of alternative splicing and gene expression data allowed to make some interesting findings. Firstly, one main principal component common to gene expression and alternative splicing events, all combined and separated by type, is clearly linked to malignancy. Indeed, samples of grades 2 and 4 glioma cases separated well away from each other, while grade 3 samples mapped along that axis, which presumably reflected a gradient of tumour aggressiveness, later corroborated by survival analyses. In contrast, LGm groups 4 and 5 could not be resolved either by gene expression or by alternative splicing main principal components of variance. Samples from LGm groups 1 and 6 showed no trend in the way they spread along the "malignancy" principal components. This behaviour is likely linked with the heterogeneity of these subtypes and was also visible in other analyses, for example while inspecting individual alternative splicing event PSI distributions. Indeed, LGm1 samples dispersion along the malignancy PC could be thought of as corresponding to various stages of disease progression from a previous, more homogeneous, LGm2 stage. In turn, LGm6 samples, which in the cohort in study corresponded at almost equal frequencies to grades 2, 3 and 4, could easily be understood to occupy the whole range of positions along gene expression and alternative splicing malignancy-associated principal components.

Glioma classifiers built based on gene expression data (i.e. transcriptome subtype and RNA expression cluster) had samples, as expected, sharply separated across the two first principal component of gene expression, while alternative splicing data provided a similar separation along combinations of principal component 2 with other main principal components. However, it was apparent from this exploratory analysis methodology plots how the high malignancy samples formed more homogeneous groups, whose diversity is not even addressed in the pan-glioma RNA-expression cluster classification.

PCA analysis of data different alternative splicing event types brought as main interesting finding that alternative 3' splice site events behaved quite differently in glioblastoma as compared to low-grade glioma. This observation actually constituted the only instance of separation of sample categories into discrete clusters coming from alternative splicing data and should be further investigated. Up to this moment, the only preliminary analysis done to try to understand the nature of this particular A3 alternative splicing behaviour was to check if A3 PSIs were more sensitive to levels of cognate gene expression, through comparative correlation analyses. This possibility was not supported. Also, the hypothesis that only a subset of A3 alternative splicing events with particular features, such as consistent splice site strength differences between the two alternative splice sites under selection, was driving the separation between GBM and LGG was briefly explored. Specifically, the distribution of loadings along PC2 of A3 alternative splicing events was inspected and, from the 2093 A3 alternative splicing events analysed, about 1248 were found to make a larger contribution to PC2. An interesting class of concurrent 3' splice sites is the one of tandem alternative splice sites (TASS), also termed NAGNAG, which constitute a particular group of A3 alternative splicing in which the mature mRNA isoforms differ by three NAG nucleotides. Although NAGNAG motifs are frequent in the human genome, only a subset of these (around 215) are subject to alternative splicing (Akerman, David-Eden, Pinter, & Mandel-Gutfreund, 2009). Because TASS alternative splicing appears to be heavily regulated and, very interestingly, switches in TASS splice site usage have been observed in cells that grow under confluence in cell culture (Szafranski et al., 2014), a condition that could be similar to the one happening in rapid growth tumour masses, the possibility of enrichment of TASS among the A3 splice sites contributing with higher variance to A3 PC2 should be explored.

While alternative splicing appeared to have similar discriminatory power towards known glioma clinical and molecular classes, it was also shown here to underlie different levels of biological information. Indeed, alternative splicing was found to keep very similar principal components of

variance when analysed without the pool of events significantly dependent on gene expression (Figure 3.9 of Results section). When functional analysis was performed on the gene expression and alternative splicing malignancy dimensions through GSEA, the enriched functional pathways found for the former were not found for the latter, except for the KEGG dilated cardiomyopathy pathway. Still, enrichment for this functional category was driven by different genes in the two analyses.

The study followed having as a main focus the elucidation of the differences in gene expression and alternative splicing regulation between LGm subtypes, which being molecularly more homogeneous than tumour grades and constituting a classification system with very valuable information in terms of prognosis, were thought of as promising entities to analyse.

A group of 5970 genes were differentially expressed between LGm groups, from which 183 corresponded to known tumour drivers and various to be associated with glioma, like ErbB tyrosine kinase receptor genes, cell proliferation related genes such as CDKN2C or MDM2, and the DNA CpG demethylator enzyme encoding *TET1*. Importantly, 41 splicing factors also appeared differentially expressed between LGm subtypes, with LGm groups 4 and 5, on the one hand, and LGm groups 2 and 3, on the other, usually presenting closer values of expression, while LGm groups 1 and 6 expressed these genes sometimes more similarly to the referred *IDH*-wild type subtypes and other times more similarly to *IDH*-mutant subtypes LGm2 and LGm3. *PTBP1*, whose overexpression in glioblastoma is known to enhance tumour survival and invasiveness, did not surpass the fold-change threshold used for selection of differential gene expression, but was considerably upregulated in LGm groups 4 and 5. Because some of its known alternative splicing event targets, such as the skipping exon 3 of *RTN4* gene, also appeared differentially spliced between LGm groups, it is possible that even small changes in PTBP1 transcriptional output are enough to switch isoform ratios of its targets in the cell. It remains to be ascertained though if this transcription factor is significantly upregulated in all glioma subtypes relatively to healthy adult brain tissues. Finally, there were 13 splicing factors among the differentially expressed genes whose RNA binding motif was known and that thus constituted good putative regulators of key glioma-specific alternative splicing events.

Analysis of differential splicing across DNA-methylation cluster subtypes was carried out using the non-parametric ANOVA equivalent Kruskal-Wallis statistical test. However, considering the usual levels of significance applied to statistical testing (e.g. FDR = 0.01 or FDR = 0.001), many differentially regulated alternative splicing events presented PSI distributions for the different LGm subtypes that overlapped in great extent. This test alone was therefore not powerful enough to efficiently identify a set of alternative splicing events reflecting clear modes of regulation particular of the different LGm subtypes, as desired. For this reason, selection of differentially regulated alternative splicing events was carried out using a significance level of 1×10^{-9} and establishing a minimum difference of median PSI values between at least two LGm groups of 0.1. A total of 1762 alternative splicing events were found to be subjected to differential splicing according to these criteria. Among the genes differentially spliced there were 89 known cancer drivers, 64 of which were not differentially expressed between glioma subtypes. This observation was quite interesting because it raised the possibility that some of these alternative splicing events might be undergoing isoform ratio changes due to somatic mutations affecting their splice sites and/or splicing regulatory elements. Although few DNA lesions present in tumours attain such high incidences that they become common traits transversal to most samples from a given tumour subtype, this hypothesis of the existence of recurrent (frequent) splicing-affecting mutations should be further analysed.

There were 46 splicing factor genes being subjected to differential splicing, none of them differentially expressed. Many of the alternative splicing events happening on these 46 genes had unknown consequences in terms of functional impact on the resulting protein. It could be interesting

to test experimentally the effect of these different splicing isoforms in cell tumourigenicity. In addition, it should be investigated which of the differentially regulated alternative splicing events involved introduction of premature stop codons and thereby putative induction of nonsense-mediated decay, so as to evaluate the extent at which splicing alterations regulate expression of certain genes in glioma cells.

Functional enrichment analysis performed on information from gene expression and alternative splicing differential regulation in LGm subtypes revealed once again differing biological functions being affected by each transcriptional process. Whereas genes with differences in expression across DNA-methylation clusters were related with functions like immune response, cell proliferation, cell survival or cell adhesion, genes having their alternative splicing affected were mostly involved in RNA-processing, protein synthesis and also apoptosis.

Although alternative splicing events displaying differential regulation across LGm groups did not seem to be able to distinguish between LGm4 and LGm5 groups, they could potentially be used for identification of LGm groups 2,3, and 4/5. This could be particularly useful for the diagnosis of glioma samples for which DNA-methylation data would not be available. LGm1 and LGm6 subtypes, shown to add important prognostic value to the DNA-methylation pan-glioma classifier published by Ceccarelli and collaborators, could not be accurately identified merely based on alternative splicing nor gene expression data. Still, it is likely that the identified glioma subtype alternative splicing markers combined with epigenetic data could help in further stratifying patients in terms of prognosis. This approach of combining epigenetic and transcriptomic data was actually already used in the lastly cited article. Specifically, grade classification assessed through histological analysis was replaced by expression of certain genes ("EReg" genes) in order to separate LGm6-LGG patients from LGm6-GBM patients.

A study of the value of alternative splicing and gene expression in glioma prognosis was then performed. Initially, a confirmation that the previously identified principal components of variance of alternative splicing and gene expression contained valuable information associated with malignancy was made. Indeed, Cox proportional-hazards models applied to explain patients' overall survival with WHO grade categories, on the one hand, and with PC loadings on the other, showed these dimensions to be advantageous over grade in glioma prognosis prediction, as judged from the Harrell's concordance index metric for evaluation of the survival regression models.

Then, the prognostic value of individual genes and alternative splicing events was evaluated, alone or after adjustment for known clinical and molecular prognostic factors. Importantly, individual alternative splicing and expressed gene markers were shown to add only negligible prognosis information to a model that already took into account LGm subtype, tumour grade and age of the patient. In fact, only two alternative splicing events and one expressed gene made a significant contribution to the previously referred multivariate Cox regression model. The two alternative splicing events in question had very narrow PSI distributions, being difficult to work with if ever used as prognostic markers and were also likely not interesting *per se* in terms of representing isoform switches with biological impact. It should be added though that it might be possible to derive a selected group of alternative splicing markers and/or genes able to add prognostic value to the glioma Cox regression model for patient's overall survival composed of LGm subtype, grade and age. The selection of markers to build such meta feature could be made from a list of prognostic markers identified in Cox regression models adjusted for LGm group and age, but not for grade. Indeed, both gene expression and alternative splicing had been shown to be superior in relation to grade in what concerns capturing malignancy.

It was then considered relevant to identify expressed genes, namely splicing factor genes, and alternative splicing events associated with LGm groups. This analysis involved the identification of markers whose prognostic values were independent of tumour grade and patient's age, since it was known that any prognostic information added to these two factors would likely be included in the LGm classifier. There were 3727 expressed genes and 237 alternative splicing events that showed to add prognostic value to the above-mentioned model settings. Further application of Cox regression to the LGG, but not to the GBM cohort separately, proved very useful, having helped to identify 493 additional glioma prognostic alternative splicing events. Unfortunately, the GBM cohort returned no significant prognostic markers able to discriminate subsets of glioblastoma patients with differential expected outcome. This is likely due to the reduced size of GBM patients used in this study, which could apparently not represent enough variation in dependent and/or independent variables in the models created in order to reach statistical significance during prognostic marker evaluation.

Finally, having in mind the aim to identify a final list of alternative splicing regulators and events clearly associated with the LGm groups and having prognostic value, the markers obtained from the survival analysis that also corresponded to differentially regulated genes and alternative splicing events were extracted. This selection returned 20 splicing factors, from which six with known RNA-binding motif, whose potential as regulators of differential alternative splicing across LGm groups was evaluated at a later stage. In turn, 337 alternative splicing events associated with DNA-methylation subtype classification were identified. From these, there were 50 that were independent of gene expression and, consistently, were also not related with genes having themselves prognostic value in glioma.

In the final results section of this manuscript, potential mechanisms of alternative splicing regulation in *trans* relevant in particular LGm glioma subtypes were looked for. Different publications dedicated to the study of alternative splicing regulation in cancer have brought evidence to the fact that alterations in exon-inclusion ratios in this disease are only rarely strongly associated with mutations in *cis* elements (Sebestyén, Zawisza, et al., 2015) or with mutations in RNA-binding protein genes (Sebestyén, Singh, et al., 2015). These two observations make it particularly pertinent to try to assess mechanisms of alternative splicing regulation through identification of strong relations between active RNA-binding splicing factors and exon-inclusion ratios of their potential targets. An attempt to find these relations was made that consisted of looking for non-random association between two variables: correlation of splicing factor gene expression with alternative splicing event quantifications and occurrence of splicing factor binding motifs in regulatory regions of alternative splicing events. This analysis was made individually for eight distinct regulatory regions of general exon-skipping events, shown in the literature to create independent contexts for context-specific alternative splicing regulation.

The likeliness for each of the six RNA-binding splicing factors RBM47, RBMS1, IGF2BP2, IGFBP3, KHDRBS2 and PCBP3 to constitute splicing regulators of a considerable portion of exon-skipping events was assessed by looking at concordance of RBP expression/PSI correlations between the GTEx and the glioma TCGA datasets. KHDRBS2 encoding gene seemed to show the best concordance between both datasets. Subsequently, information about the actual presence of binding motifs for these RBPs in SE events regulatory regions was collected, promisingly showing an enrichment in the number of alternative splicing events that contained RBP motifs (*cis* elements) together with significant correlation with RBP expression among those that were differentially regulated across LGm subtypes.

RNA splicing maps were then derived, first for PTBP1, in order to validate the algorithm used, and then to the other RBPs with strong association with LGm subtypes. RNA splicing maps for PTBP1 very

nicely reproduced what had been found in the literature in terms of the mode of action of this splicing regulator, both using GTEx multi tissue and TCGA glioma datasets. However, for the other RBPs, discrepancies were found between the maps produced using both datasets. Still, the KHDRBS2 GTEx-derived RNA splicing maps seemed quite reliable, both in terms of the level of significance of the highest peaks of discovered regulatory regions and also the stability of their relative position under different parameter settings (data not shown).

In the future, some improvements to the RNA splicing map generating algorithm used here may be made in order to be able to take stronger conclusions on RBP-specific alternative splicing regulation mechanisms and, in particular, about the possibility to detect common modes of regulation in tumourigenic *vs* healthy tissue. As such, the main proposed change on the algorithm itself would be to start accounting for the total number of RBP motifs in each region searched. In our analyses, enrichment tests were run on the binary information of each alternative splicing event being bound or not, meaning a maximum count of one motif per event. Then, TCGA data should be analysed using a more comprehensive transcriptomic annotation, so that more alternative splicing events can be profiled and the power of the statistical tests thereby increased.

Some concluding remarks may be noted. The work presented here allowed the elucidation of the relative contribution of alternative splicing for glioma subtype assessment, namely relatively to gene expression and DNA-methylation data levels. Characterization of alternative splicing profiles in different glioma grades and LGm subtypes brought some interesting new findings that can be further explored to help in the identification of mechanisms of splicing regulation affected in particular glioma groups and also in the improvement of patient clinical management, following to a better stratification according to prognosis. As main drawbacks associated with this work, limiting the discovery potential but not compromising the quality of the results presented, were the sub-optimal transcriptome annotation underlying the glioma data and impossibility of profiling tumour-specific aberrant splicing due precisely to the use of data processed based on a reference transcriptome.

5 REFERENCES

Akerman, M., David-Eden, H., Pinter, R. Y., & Mandel-Gutfreund, Y. (2009). A computational approach for genome-wide mapping of splicing factor binding sites. *Genome Biology*, *10*, R30. JOUR. http://doi.org/10.1186/gb-2009-10-3-r30

Alamancos, G. P., Pagès, A., Trincado, J. L., Bellora, N., & Eyras, E. (2014). SUPPA: a super-fast pipeline for alternative splicing analysis from RNA-Seq. *bioRxiv*, 8763. http://doi.org/10.1101/008763

Anaya, J., Reon, B., Chen, W.-M., Bekiranov, S., & Dutta, A. (2016). A pan-cancer analysis of prognostic genes. *PeerJ*, *3*. JOUR. http://doi.org/10.7717/peerj.1499

Annayev, Y., Adar, S., Chiou, Y.-Y., Lieb, J. D., Sancar, A., & Ye, R. (2014). Gene model 129 (Gm129) encodes a novel transcriptional repressor that modulates circadian gene expression. *The Journal of Biological Chemistry*, *289*(8), 5013–5024. JOUR. http://doi.org/10.1074/jbc.M113.534651

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. JOUR. Retrieved from http://www.jstor.org/stable/2346101

Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., … McLendon, R. (2013). The somatic genomic landscape of glioblastoma. *Cell*, *155*(2), 462–477. http://doi.org/10.1016/j.cell.2013.09.034

cBioPortal for Cancer Genomics. (2016, August 22). ELEC. Retrieved from http://www.cbioportal.org/

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., … Verhaak, R. G. W. (2016). Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, *164*(3), 550–563. http://doi.org/10.1016/j.cell.2015.12.028

Chen, C.-Y., Logan, R. W., Ma, T., Lewis, D. A., Tseng, G. C., Sibille, E., & McClung, C. A. (2016). Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(1), 206–211. JOUR. http://doi.org/10.1073/pnas.1508249112

Cheung, H. C., Hai, T., Zhu, W., Baggerly, K. A., Tsavachidis, S., Krahe, R., & Cote, G. J. (2009). Splicing factors PTBP1 and PTBP2 promote proliferation and migration of glioma cell lines. *Brain: A Journal of Neurology*, *132*(Pt 8), 2277–2288. JOUR. http://doi.org/10.1093/brain/awp153

Cleynen, I., Brants, J. R., Peeters, K., Deckers, R., Debiec-Rychter, M., Sciot, R., … Petit, M. M. R. (2007). HMGA2 regulates transcription of the Imp2 gene via an intronic regulatory element in cooperation with nuclear factor-kappaB. *Molecular Cancer Research: MCR*, *5*(4), 363–372. JOUR. http://doi.org/10.1158/1541-7786.MCR-06-0331

COSMIC: Catalogue of Somatic Mutations in Cancer - Home Page. (2016, August 22). ELEC. Retrieved from http://cancer.sanger.ac.uk/cosmic

Danan-Gotthold, M., Golan-Gerstl, R., Eisenberg, E., Meir, K., Karni, R., & Levanon, E. Y. (2015). Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic Acids Research*, *43*(10), 1–15. http://doi.org/10.1093/nar/gkv210

de la Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., … Kornblihtt, A. R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. *Molecular Cell*, *12*(2), 525–

532. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14536091

Dujardin, G., Lafaille, C., de la Mata, M., Marasco, L. E., Muñoz, M. J., Le Jossic-Corcos, C., … Kornblihtt, A. R. (2014). How slow RNA polymerase II elongation favors alternative exon skipping. *Molecular Cell*, *54*(4), 683–690. JOUR. http://doi.org/10.1016/j.molcel.2014.03.044

edgeR. (2016, September 25). *Bioconductor*. ELEC. Retrieved from http://bioconductor.org/packages/edgeR/

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. (2016, September 1). ELEC. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/

Ferrarese, R., Harsh, G. R., Yadav, A. K., Bug, E., Maticzka, D., Reichardt, W., … Bredel, M. (2014). Lineage-specific splicing of a brain-enriched alternative exon promotes glioblastoma progression. *The Journal of Clinical Investigation*, *124*(7), 2861–2876. JOUR. http://doi.org/10.1172/JCI68836

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., … Campbell, P. J. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, *43*(D1), D805–D811. JOUR. http://doi.org/10.1093/nar/gku1075

Fundamentals of Biostatistics 7th edition (9780538733496) - Textbooks.com. (2016, August 25). ELEC. Retrieved from http://www.textbooks.com/Fundamentals-of-Biostatistics-7th-Edition/9780538733496/Bernard-Rosner.php

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., … Stratton, M. R. (2004). A CENSUS OF HUMAN CANCER GENES. *Nature Reviews. Cancer*, *4*(3), 177–183. JOUR. http://doi.org/10.1038/nrc1299

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, *27*(7), 1017–1018. JOUR. http://doi.org/10.1093/bioinformatics/btr064

Graphics and Data Visualization in R. (2016, September 14). ELEC. Retrieved from http://girke.bioinformatics.ucr.edu/GEN242/vignettes/15_Rgraphics/Rgraphics.html

GTEx Portal. (2016, August 24). ELEC. Retrieved from http://www.gtexportal.org/home/

GTEx Quantifications - Flux Capacitor - Confluence. (2016, August 31). ELEC. Retrieved from http://sammeth.net/confluence/display/FLUX/GTEx+Quantifications

Hall, M. P., Nagel, R. J., Fagg, W. S., Shiue, L., Cline, M. S., Perriman, R. J., … Ares, M. (2013). Quaking and PTB control overlapping splicing regulatory networks during muscle cell differentiation. *RNA*, *19*(5), 627–638. JOUR. http://doi.org/10.1261/rna.038422.113

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, *247*(18), 2543–2546. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7069920

Hu, J., Ho, A. L., Yuan, L., Hu, B., Hua, S., Hwang, S. S., … DePinho, R. A. (2013). From the Cover: Neutralization of terminal differentiation in gliomagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(36), 14520–14527. JOUR. http://doi.org/10.1073/pnas.1308610110

Karni, R., de Stanchina, E., Lowe, S. W., Sinha, R., Mu, D., & Krainer, A. R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nature Structural & Molecular Biology*, *14*(3), 185–193. JOUR. http://doi.org/10.1038/nsmb1209

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. JOUR. http://doi.org/10.1186/1471-2105-12-323

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500. JOUR. http://doi.org/10.1093/bioinformatics/btp692

Li, Y. I., Sanchez-Pulido, L., Haerty, W., & Ponting, C. P. (2015). RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Research*, *25*(1), 1–13. JOUR. http://doi.org/10.1101/gr.181990.114

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., … Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, *456*(7221), 464–469. JOUR. http://doi.org/10.1038/nature07488

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. JOUR. http://doi.org/10.1038/ng.2653

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., … Kleihues, P. (2007). The 2007 WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathologica*, *114*(2), 97–109. JOUR. http://doi.org/10.1007/s00401-007-0243-4

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., … Ellison, D. W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica*, *131*(6), 803–820. JOUR. http://doi.org/10.1007/s00401-016-1545-1

Matera, A. G., & Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*, *15*(2), 108–121. JOUR. http://doi.org/10.1038/nrm3742

Modrek, A. S., Bayin, N. S., & Placantonakis, D. G. (2014). Brain stem cells as the cell of origin in glioma. *World Journal of Stem Cells*, *6*(1), 43–52. JOUR. http://doi.org/10.4252/wjsc.v6.i1.43

Moehle, E. A., Braberg, H., Krogan, N. J., & Guthrie, C. (2014). Adventures in time and space. *RNA Biology*, *11*(4), 313–319. JOUR. http://doi.org/10.4161/rna.28646

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., … Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, *464*(7289), 773–777. http://doi.org/10.1038/nature08903.Transcriptome

Munro, T. P., Magee, R. J., Kidd, G. J., Carson, J. H., Barbarese, E., Smith, L. M., & Smith, R. (1999). Mutational analysis of a heterogeneous nuclear ribonucleoprotein A2 response element for RNA trafficking. *The Journal of Biological Chemistry*, *274*(48), 34389–34395. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10567417

Nedergaard, M., Ransom, B., & Goldman, S. A. (2016). New roles for astrocytes: Redefining the functional architecture of the brain. *Trends in Neurosciences*, *26*(10), 523–530. JOUR. http://doi.org/10.1016/j.tins.2003.08.008

Network, T. C. G. A. (TCGA) R. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068. JOUR. http://doi.org/10.1038/nature07385

Nichols, R. C., Wang, X. W., Tang, J., Hamilton, B. J., High, F. A., Herschman, H. R., & Rigby, W. F.

(2000). The RGG domain in hnRNP A2 affects subcellular localization. *Experimental Cell Research*, *256*(2), 522–532. JOUR. http://doi.org/10.1006/excr.2000.4827

Nielsen, J., Christiansen, J., Lykke-Andersen, J., Johnsen, A. H., Wewer, U. M., & Nielsen, F. C. (1999). A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. *Molecular and Cellular Biology*, *19*(2), 1262–1270. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9891060

Nishimura, Y., Komatsu, S., Ichikawa, D., Nagata, H., Hirajima, S., Takeshita, H., … Otsuji, E. (2013). Overexpression of YWHAZ relates to tumor cell proliferation and malignant outcome of gastric carcinoma. *British Journal of Cancer*, *108*(6), 1324–1331. JOUR. http://doi.org/10.1038/bjc.2013.65

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., … Network, C. G. A. R. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, *17*(5), 510–522. JOUR. http://doi.org/10.1016/j.ccr.2010.03.017

Ostrom, Q. T., Bauchet, L., Davis, F. G., Deltour, I., Fisher, J. L., Langer, C. E., … Barnholtz-Sloan, J. S. (2014). The epidemiology of glioma in adults: A state of the science review. *Neuro-Oncology*, *16*(7), 896–913. http://doi.org/10.1093/neuonc/nou087

Pacheco, T. R., Gomes, A. Q., Barbosa-Morais, N. L., Benes, V., Ansorge, W., Wollerton, M., … Carmo-Fonseca, M. (2004). Diversity of vertebrate splicing factor U2AF35: identification of alternatively spliced U2AF1 mRNAS. *The Journal of Biological Chemistry*, *279*(26), 27039–27049. JOUR. http://doi.org/10.1074/jbc.M402136200

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–1415. JOUR. http://doi.org/10.1038/ng.259

Park, J. W., Jung, S., Rouchka, E. C., Tseng, Y.-T., & Xing, Y. (2016). rMAPS: RNA map analysis and plotting server for alternative exon regulation. *Nucleic Acids Research*, gkw410. http://doi.org/10.1093/nar/gkw410

Paz, I., Kosti, I., Ares, M., Cline, M., & Mandel-Gutfreund, Y. (2014). RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Research*, gku406. JOUR. http://doi.org/10.1093/nar/gku406

Percentage Splicing Index - Geuvadis MediaWiki. (2016, September 3). ELEC. Retrieved from http://geuvadiswiki.crg.es/index.php/Percentage_Splicing_Index

Prentice, R. L. (1992). Introduction to Cox (1972) Regression Models and Life-Tables. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 519–526). CHAP, New York, NY: Springer New York. Retrieved from http://dx.doi.org/10.1007/978-1-4612-4380-9_36

Przychodzen, B., Jerez, A., Guinta, K., Sekeres, M. A., Padgett, R., Maciejewski, J. P., & Makishima, H. (2013). Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. *Blood*, *122*(6), 999–1006. JOUR. http://doi.org/10.1182/blood-2013-01-480970

Raj, B., & Blencowe, B. J. (2015). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron*, *87*(1), 14–27. http://doi.org/10.1016/j.neuron.2015.05.004

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., … Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, *499*(7457), 172–177.

JOUR. http://doi.org/10.1038/nature12311

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. JOUR. http://doi.org/10.1093/nar/gkv007

RNASeq Version 2 - TCGA - National Cancer Institute - Confluence Wiki. (2016, August 31). ELEC. Retrieved from https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. JOUR. http://doi.org/10.1186/gb-2010-11-3-r25

Rosner, B. (Bernard A. . (2011). *Fundamentals of biostatistics*. BOOK, Seventh edition. Boston : Brooks/Cole, Cengage Learning, [2011] ©2011. Retrieved from https://search.library.wisc.edu/catalog/9910098055302121

Schaeffer, D. F., Owen, D. R., Lim, H. J., Buczkowski, A. K., Chung, S. W., Scudamore, C. H., … Owen, D. A. (2010). Insulin-like growth factor 2 mRNA binding protein 3 (IGF2BP3) overexpression in pancreatic ductal adenocarcinoma correlates with poor survival. *BMC Cancer*, *10*, 59. JOUR. http://doi.org/10.1186/1471-2407-10-59

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M. A., … Eyras, E. (2015). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *bioRxiv*, 23010. http://doi.org/10.1101/023010

Sebestyén, E., Zawisza, M., & Eyras, E. (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Research*, *43*(3), 1345–1356. JOUR. http://doi.org/10.1093/nar/gku1392

Snell, R. S. (2010). *Clinical Neuroanatomy*. BOOK, Wolters Kluwer Health/Lippincott Williams & Wilkins. Retrieved from https://books.google.pt/books?id=ABPmvroyrD0C

Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D. A., Jones, D. T. W., Konermann, C., … Pfister, S. M. (2012). Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. *Cancer Cell*, *22*(4), 425–437. http://doi.org/10.1016/j.ccr.2012.08.024

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. JOUR. http://doi.org/10.1073/pnas.0506580102

Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., … Ogawa, S. (2015). Mutational landscape and clonal architecture in grade II and III gliomas. *Nat Genet*, *47*(5), 458–468. http://doi.org/10.1038/ng.3273

Szafranski, K., Fritsch, C., Schumann, F., Siebel, L., Sinha, R., Hampe, J., … Platzer, M. (2014). Physiological state co-regulates thousands of mammalian mRNA splicing events at tandem splice sites and alternative exons. *Nucleic Acids Research*, *42*(14), 8895–8904. JOUR. http://doi.org/10.1093/nar/gku532

TCGA Home - TCGA - National Cancer Institute - Confluence Wiki. (2016, August 22). ELEC. Retrieved from https://wiki.nci.nih.gov/display/TCGA/TCGA+Home

The Cancer Genome Atlas - Data Portal. (2016, August 22). ELEC. Retrieved from https://tcga-data.nci.nih.gov/docs/publications/tcga/

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, *99*(10), 6567–6572. JOUR. http://doi.org/10.1073/pnas.082099299

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. JOUR. http://doi.org/10.1093/bioinformatics/btp120

Trombetta-Lima, M., Winnischofer, S. M. B., Demasi, M. A. A., Filho, R. A., Carreira, A. C. O., Wei, B., … Sogayar, M. C. (2015). Isolation and characterization of novel RECK tumor suppressor gene splice variants. *Oncotarget*, *6*(32), 33120–33133. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4741753/

Tsai, Y. S., Dominguez, D., Gomez, S. M., & Wang, Z. (2015). Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors. *Oncotarget*, *6*(9), 1–15. http://doi.org/10.18632/oncotarget.3145

Venables, J. P., Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Koh, C., … Elela, S. A. (2008). Identification of Alternative Splicing Markers for Breast Cancer. *Cancer Research*, *68*(22), 9525–9531. JOUR. http://doi.org/10.1158/0008-5472.CAN-08-1769

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., … Hayes, D. N. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, *17*(1), 98–110. JOUR. http://doi.org/10.1016/j.ccr.2009.12.020

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., … Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. JOUR. http://doi.org/10.1038/nature07509

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., … Liu, J. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, *38*(18), e178. JOUR. http://doi.org/10.1093/nar/gkq622

Wang, Z.-L. Z., Zhang, C.-B., Cai, J.-Q., Li, Q.-B., Wang, Z.-L. Z., & Jiang, T. (2015). Integrated analysis of genome-wide DNA methylation, gene expression and protein expression profiles in molecular subtypes of WHO II-IV gliomas. *Journal of Experimental & Clinical Cancer Research : CR*, *34*, 127. http://doi.org/10.1186/s13046-015-0249-z

Weyn-Vanhentenryck, S. M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., … Zhang, C. (2014). HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Reports*, *6*(6), 1139–1152. JOUR. http://doi.org/10.1016/j.celrep.2014.02.005

Yamada, S. M., Yamaguchi, F., Brown, R., Berger, M. S., & Morrison, R. S. (1999). Suppression of glioblastoma cell growth following antisense oligonucleotide-mediated inhibition of fibroblast growth factor receptor expression. *Glia*, *28*(1), 66–76. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10498824

Yamaguchi, F., Saya, H., Bruner, J. M., & Morrison, R. S. (1994). Differential expression of two fibroblast growth factor-receptor genes is associated with malignant progression in human astrocytomas. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(2), 484–488. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC42973/

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., … Vidal, M. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*,

164(4), 805–817. http://doi.org/10.1016/j.cell.2016.01.029

Yoon, J.-H., De, S., Srikantan, S., Abdelmohsen, K., Grammatikakis, I., Kim, J., … Gorospe, M. (2014). PAR-CLIP analysis uncovers AUF1 impact on target RNA fate and genome integrity. *Nature Communications*, *5*, 5248. JOUR. http://doi.org/10.1038/ncomms6248

You, F., Sun, H., Zhou, X., Sun, W., Liang, S., Zhai, Z., & Jiang, Z. (2009). PCBP2 mediates degradation of the adaptor MAVS via the HECT ubiquitin ligase AIP4. *Nature Immunology*, *10*(12), 1300–1308. JOUR. http://doi.org/10.1038/ni.1815

Zhang, J. Y., Chan, E. K., Peng, X. X., & Tan, E. M. (1999). A novel cytoplasmic protein with RNA-binding motifs is an autoantigen in human hepatocellular carcinoma. *The Journal of Experimental Medicine*, *189*(7), 1101–1110. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10190901

Zwiener, I., Frisch, B., & Binder, H. (2014). Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. *PLOS ONE*, *9*(1), e85150. JOUR. http://doi.org/10.1371/journal.pone.0085150

6 SUPPLEMENTS

Table S 1 – Class designations, histological codes and tumour grading of gliomas by the 2016 WHO Classification of CNS tumours (Louis et al., 2016).

Designation	***Histology ICH id****	***WHO grade***
Diffuse astrocytic and oligodendroglial tumours		
Diffuse astrocytoma, IDH-mutant[a]	9400/3	II
Diffuse astrocytoma, IDH-wildtype[a]	9400/3	II
Anaplastic astrocytoma, IDH-mutant[a]	9401/3	III
Anaplastic astrocytoma, IDH-wildtype[a]	9401/3	III
Glioblastoma, IDH-wildtype[b]	9440/3	IV
Glioblastoma, IDH-mutant[b]	9445/3	IV
Diffuse midline glioma, H3 K27M-mutant	9385/3	IV
Oligodendroglioma, IDH-mutant and 1p/19q-codeleted[a]	9450/3	II
Anaplastic oligodendroglioma, IDH-mutant and 1p/19q-codeleted[a]	9451/3	III
Other astrocytic tumours		
Pilocytic astrocytoma	9421/3	I
Subependymal giant cell astrocytoma	9384/1	I
Pleomorphic xanthoastrocytoma	9424/3	II
Anaplastic pleomorphic xanthoastrocytoma	9424/3	III
Other gliomas		
Chordoid glioma of the third ventricle	9444/1	II
Angiocentric glioma	9431/1	I

* Morphology code coming from the International Classification of Diseases for Oncology

[a] Included in LGG-TCGA cohort

[b] Included in GBM-TCGA cohort

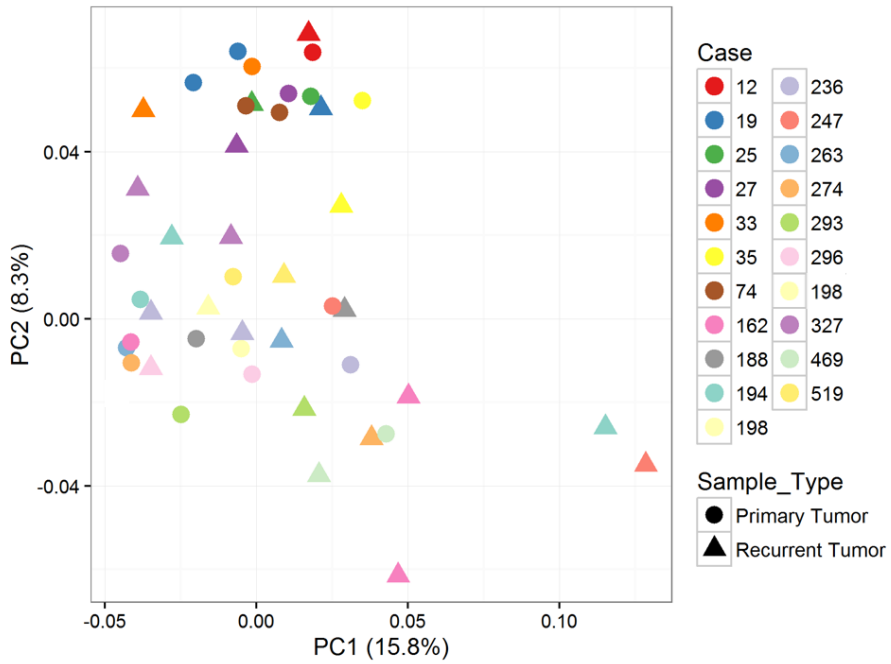


Figure S 1 – Scatter plot of primary-recurrent glioma paired samples across PSI matrix principal components 1 and 2. Samples from the same patient are identified by one "Case" colour and the plotted symbol illustrates the sample type. Underlying PCA was performed on the whole glioma cohort and only paired samples plotted for a clearer visualization of their relative positions.

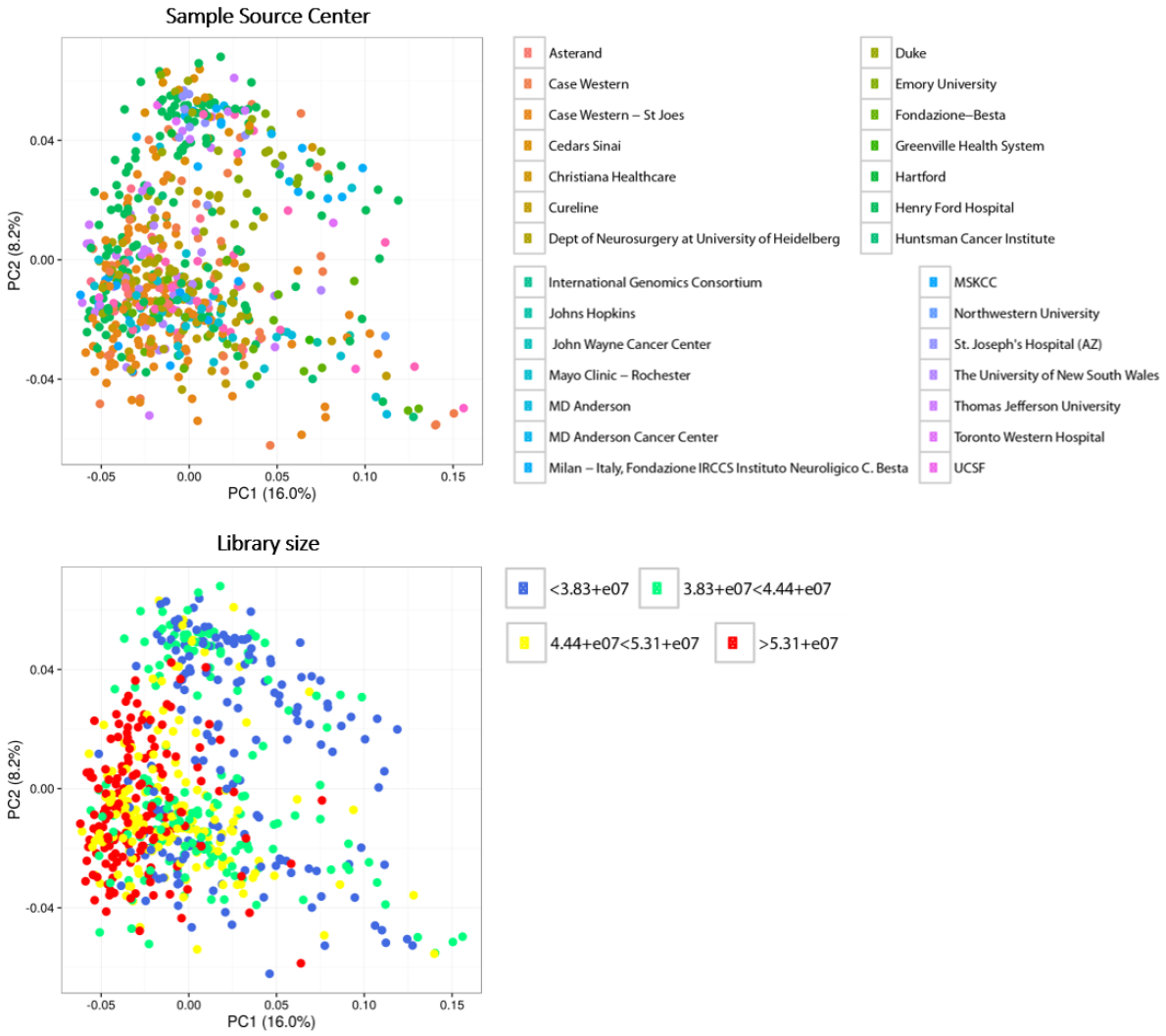


Figure S 2 – Principal Component Analysis scatter plots of alternative splicing data, coloured for sample source centre and sample library size. Numbers in the library size labels represent total numbers of mapped reads, with samples separated by quartiles of library sizes in the glioma cohort. Numbers in the library size labels represent total numbers of mapped reads and correspond to size boundaries that correspond to the three quartiles of library sizes in the glioma cohort.

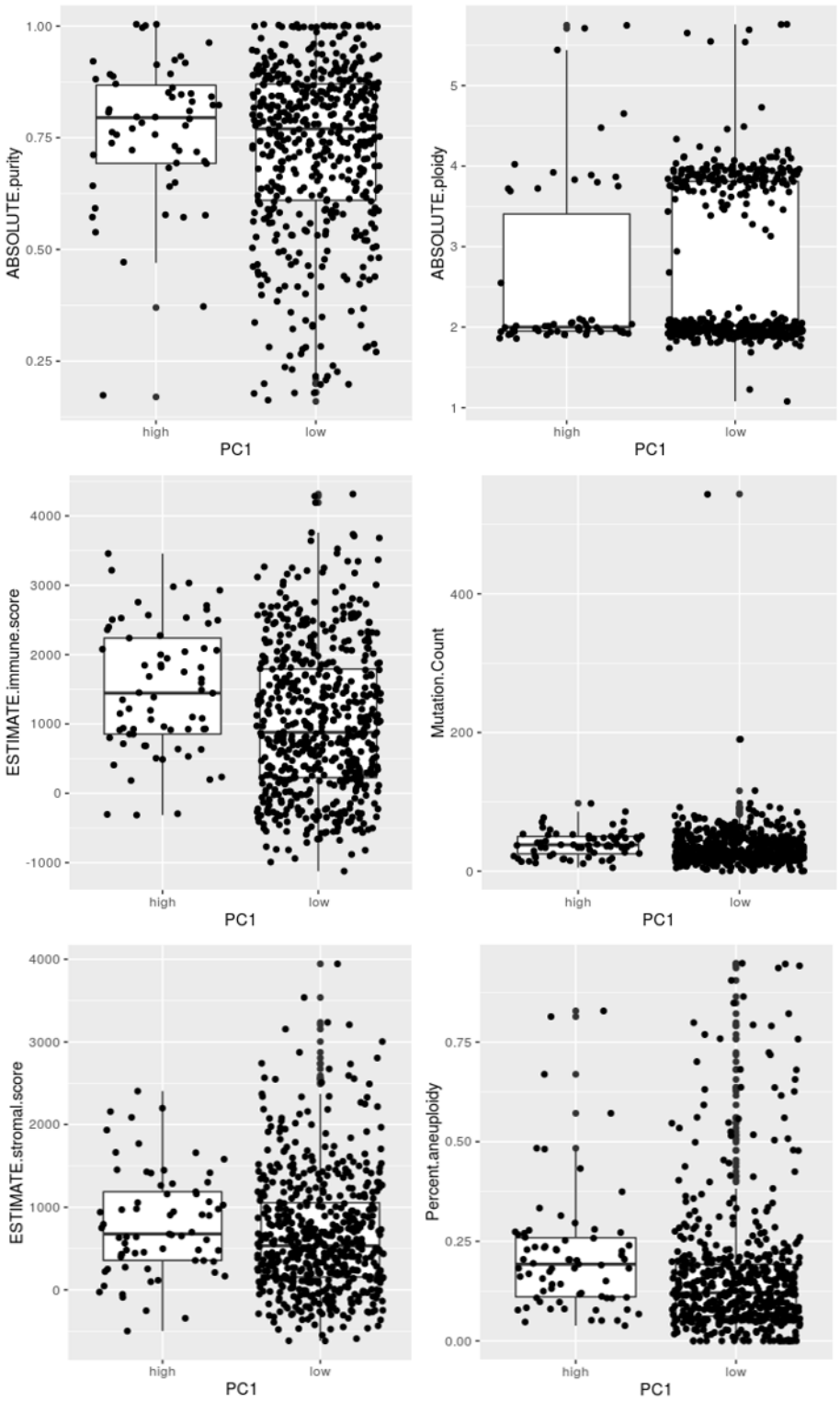


Figure S 3 – Distributions of different parameters related with somatic DNA alterations and tumour purity in two groups of samples behaving differently along alternative splicing principal component 1.

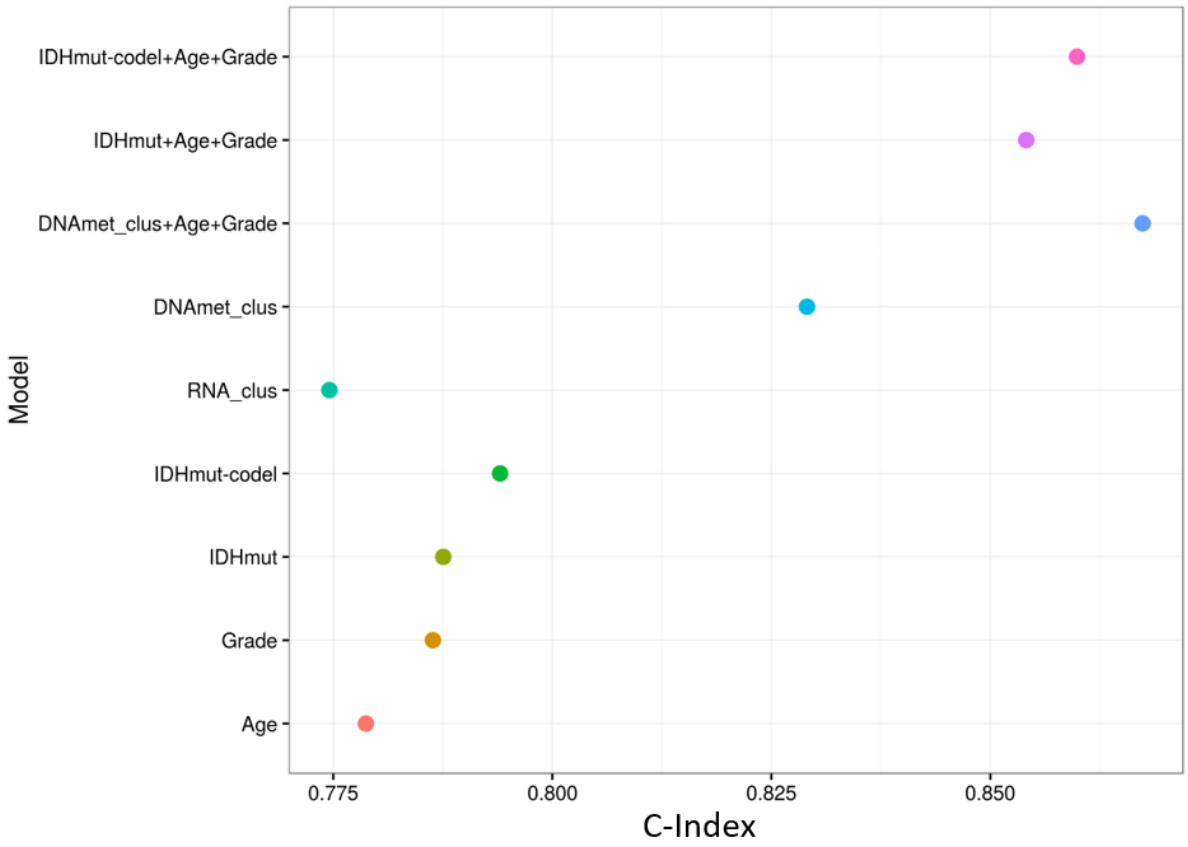


Figure S 4 – Concordance indexes for different Cox regression models applied to the glioma cohort, using recognized clinical and molecular variables.

Table S 2 – Summary statistics of Cox regression models including significant alternative splicing events, after adjustment for gene expression, DNA-methylation cluster, WHO grade and age. FDR cut-off value: 0.05.

Event	*Gene*	*GE.HR*	*PSI.HR*	*Concordance*	*PSI.FDR*	*logrank.FDR*
COPS4;SE:chr4:83989675-83994358:83994506-83996449:+	COPS4	0,78	2,30E-09	0,865	0,02	0,00
CTSB;SE:chr8:11710988-11715366:11715449-11725509:-	CTSB	1,46	7,29E+32	0,875	0,04	0,00
DENND5B;SE:chr12:31648853-31652539:31652605-31743639:-	DENND5B	1,08	4,32E+06	0,865	0,02	0,00
FGF1;SE:chr5:141993726-142077186:142077280-142077477:-	FGF1	0,95	8,99E-02	0,868	0,02	0,00
LPXN;SE:chr11:58322413-58331627:58331674-58338028:-	LPXN	1,04	1,03E+02	0,865	0,03	0,00
LRIG1;SE:chr3:66449465-66452032:66452104-66455621:-	LRIG1	0,93	2,94E+25	0,864	0,03	0,00
MIB2;SE:chr1:1560808-1560925:1561033-1562029:+	MIB2	0,72	1,20E-02	0,868	0,03	0,00
MRO;SE:chr18:48326513-48327718:48327874-48331523:-	MRO	0,92	7,46E+00	0,865	0,04	0,00
NAGPA;SE:chr16:5078186-5078297:5078399-5078880:-	NAGPA	1,26	1,90E-05	0,864	0,04	0,00
NAP1L4;SE:chr11:3000467-3010358:3010501-3013483:-	NAP1L4	0,81	3,93E+10	0,865	0,02	0,00
NLGN4X;SE:chrX:6069812-6105523:6105830-6146581:-	NLGN4X	1,03	1,99E+30	0,865	0,00	0,00
PDE8A;SE:chr15:85632647-85634274:85634412-85641178:+	PDE8A	0,98	4,94E-02	0,857	0,02	0,00
PDGFRA;SE:chr4:55124984-55127261:55127579-55129833:+	PDGFRA	1,04	5,11E-10	0,871	0,00	0,00
POLR2J4;SE:chr7:44053278-44054204:44054382-44056032:-	POLR2J4	1,02	3,84E+01	0,866	0,05	0,00
PSMB5;SE:chr14:23495584-23496953:23497038-23502576:-	PSMB5	1,17	8,17E+17	0,866	0,04	0,00
RBM42;SE:chr19:36125275-36128059:36128254-36128343:+	RBM42	0,95	5,75E-186	0,862	0,04	0,00
RMND5B;SE:chr5:177558377-177562173:177562313-177565108:+	RMND5B	0,57	7,87E-07	0,870	0,03	0,04
SLIT1;SE:chr10:98791434-98794227:98794299-98797454:-	SLIT1	0,94	1,03E-04	0,865	0,04	0,00
SNRPN;SE:chr15:25207356-25213078:25213229-25219457:+	SNRPN	0,67	3,72E-48	0,868	0,03	0,00
STYXL1;SE:chr7:75625917-75630207:75630320-75633075:-	STYXL1	1,08	1,74E-02	0,864	0,05	0,00
ZNF33A;SE:chr10:38299711-38301225:38301278-38305798:+	ZNF33A	0,64	1,45E+01	0,867	0,04	0,00
ASTN2;RI:chr9:119187506:119187905-119188188:119188367:-	ASTN2	1,00	1,33E-04	0,865	0,04	0,00
HRAS;RI:chr11:532242:532522-532630:532755:-	HRAS	0,88	1,63E+03	0,865	0,03	0,01
CWC25;A5:chr17:36977326-36981418:36977326-36981522:-	CWC25	0,75	6,86E-04	0,866	0,02	0,00
PCBP4;A5:chr3:51993308-51993378:51993308-51993382:-	PCBP4	0,94	2,32E+23	0,865	0,04	0,00
PNPLA8;A5:chr7:108154737-108154875:108154737-108154879:-	PNPLA8	0,79	4,91E+13	0,863	0,02	0,00
SAE1;A5:chr19:47634369-47646750:47634285-47646750:+	SAE1	0,84	8,76E+01	0,866	0,03	0,00
TMX1;A5:chr14:51712172-51713809:51712076-51713809:+	TMX1	1,04	1,60E+22	0,867	0,03	0,00
TOMM5;A5:chr9:37588929-37592306:37588929-37592408:-	TOMM5	1,06	1,06E+17	0,869	0,05	0,00
TTC8;A5:chr14:89307272-89307380:89307267-89307380:+	TTC8	0,75	3,62E+60	0,871	0,03	0,00
CDV3;A3:chr3:133305566-133306002:133305566-133306005:+	CDV3	1,05	6,66E+01	0,867	0,02	0,00
IMMT;A3:chr2:86398459-86400772:86398435-86400772:-	IMMT	1,34	2,51E+40	0,859	0,02	0,00
RANBP1;A3:chr22:20113891-20114474:20113891-20114477:+	RANBP1	0,85	4,25E-03	0,867	0,03	0,00
RBM8A;A3:chr1:145508075-145508206:145508075-145508209:+	RBM8A	0,98	2,16E-03	0,868	0,03	0,00
ELOVL1;AF:chr1:43831294-43833156:43833361:43831294-43833585:43833699:-	ELOVL1	1,09	2,60E+09	0,869	0,02	0,00
HNRNPUL1;AF:chr19:41770639:41770703-41774127:41771026:41771248-41774127:+	HNRNPUL1	0,88	7,09E-02	0,867	0,04	0,00
MED15;AF:chr22:20861885:20862033-20891403:20862338:20862731-20891403:+	MED15	0,69	3,22E-05	0,866	0,05	0,00
NDRG2;AF:chr14:21492255-21492984:21493185:21492255-21493835:21493935:-	NDRG2	1,01	3,86E-05	0,869	0,03	0,00
TSC22D3;AF:chrX:106959180-106959544:106959711:106959180-107018329:107019017:-	TSC22D3	0,90	3,89E-02	0,865	0,04	0,00
TUBB3;AF:chr16:89988416:89988653-89998978:89989744:89989866-89998978:+	TUBB3	1,27	1,50E+14	0,870	0,03	0,00
EIF4E2;AL:chr2:233431924-233433654:233433919:233431924-233445613:233448349:+	EIF4E2	0,66	1,03E+01	0,864	0,04	0,00

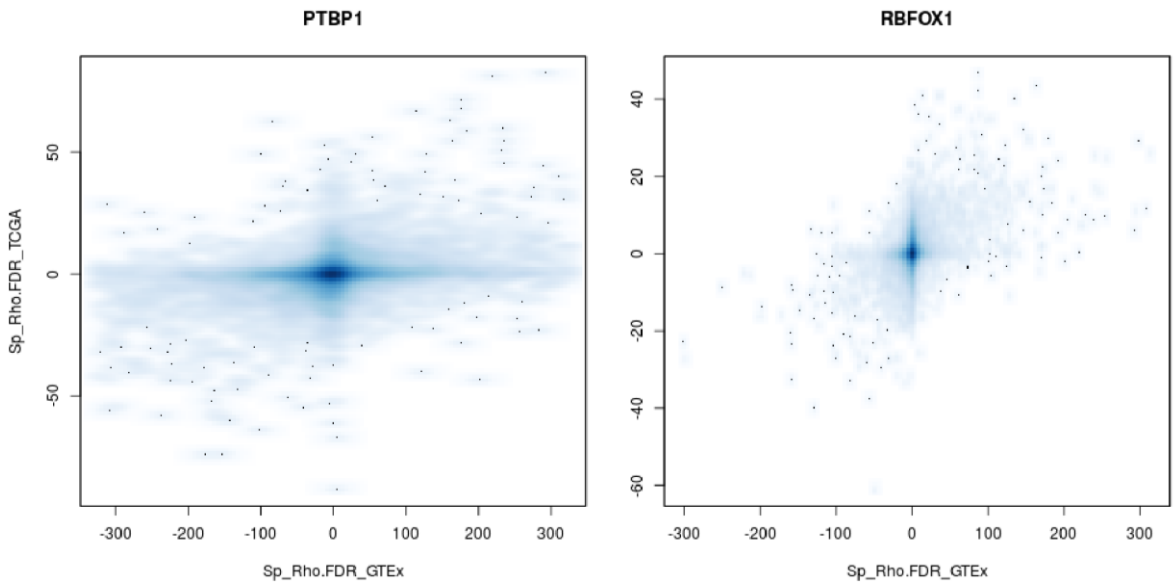


Figure S 5 – Concordance between glioma TCGA and GTEx established splicing factor to alternative splicing events PSIs correlations. Scatter plots comparing the two data sets for -log10(FDR) values times the sign of the Spearman correlation rho between RBP expression and PSIs, for each RBP splicing factor.

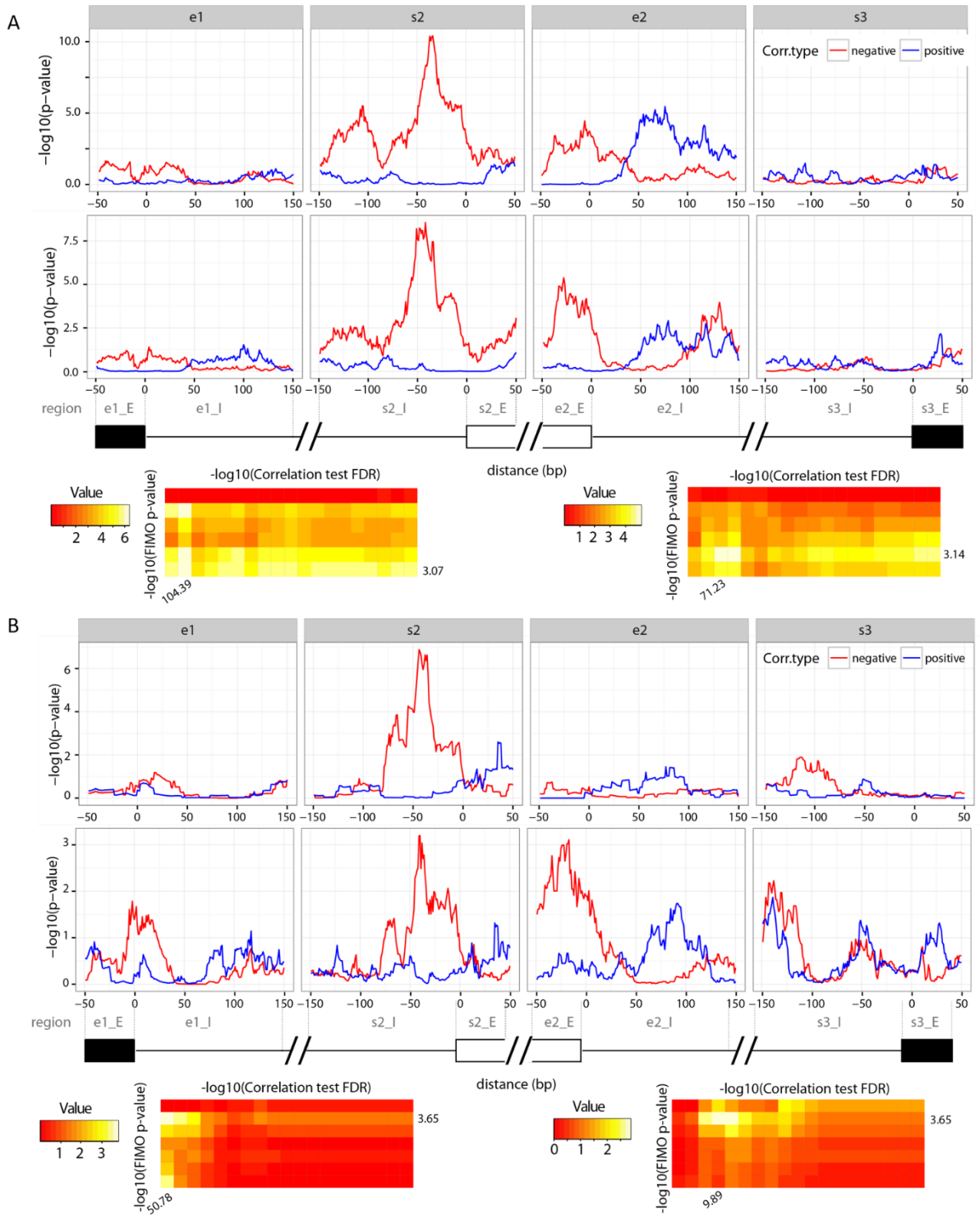


Figure S 6 – PTBP1 RNA-binding maps for the general exon-skipping (SE) alternative splicing event. . (A) Two RNA-binding maps produced using the GTEx multi-tissue dataset. (B) Two RNA-binding maps produced using the glioma TCGA dataset. RNA-binding maps shown were generated using correlation FDR threshold and FIMO p-value as shown on the bottom and the right side of the heat maps, respectively. Distance in base pairs (bp) relative to the closest splice site is shown. Different names for the eight intronic (150 nucleotides long) and exonic (50 nucleotides long) regulatory regions defined are indicated in grey. Constitutive exons are shown in black and alternative exon is shown in white. Corr-type – Correlation type: blue for enhancement and red for silencing of exon inclusion.